Modeling interpretation in the age of GenAI: semantic technologies for digital humanities research based on Giorgio Vasari's The Lives

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

von der KIT-Fakultät für
Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

M.Sc. Sarah Rebecca Ondraszek
geb. in Hildesheim

Tag der mündlichen Prüfung:	23.10.2026
Hauptreferent:	Prof. Dr. York Sure-Vetter
Korreferent:	Prof. Dr. Clemens Beckstein

Karlsruhe (2026)

To live for the hope of it all

Kurzfassung

Diese Dissertation etabliert die Hauptthese, dass maschinelle Extraktion aus interpretativen Quellen selbst ein Akt der Quellenproduktion ist und daher derselben Quellenkritik unterzogen werden muss. Sie entwickelt diesen Grundsatz zu einem formalen Rahmen für die Konstruktion von Wissensgraphen weiter, in dem jede Aussage, die ein generatives Modell extrahiert, als zu kritisierende Quelle und nicht als akzeptierte Tatsache behandelt wird.

Der zentrale wissenschaftliche Beitrag liegt in dieser Erkenntnis, dass die Konstruktion von Wissensgraphen aus interpretativen historischen Quellen die explizite Darstellung von Interpretationen und deren Herkunft als eigenständige Wissenseinheiten erfordert, um Rückverfolgbarkeit, kritische Bewertung und wissenschaftliche Transparenz zu gewährleisten.

Wenn Large Language Models Wissen aus interpretativen historischen Texten extrahieren, scheint der Output oft wie getroffene Entscheidungen, die sonst menschliche Lesende erschließen würden. So etwa, ob "der Kulturminister am Hof der Medici in Florenz" auf Giorgio Vasari verweist oder ob ein "kleines Andachtsbild" ein bestimmtes Kunstwerk bezeichnet – und welches. Konventionelle Knowledge-Graph-Konstruktions-Pipelines setzen voraus, dass solche Mention-to-Referent-Auflösungen bereits vor der Repräsentation abgeschlossen sind. Diese Dissertation stellt diese, wie wir sie bezeichnen, Pre-Interpretive Entity Assumption in Frage.

Wir betonen dies als grundlegenden Defekt bei der Erstellung von Knowledge Graphs, insbesondere bei interpretativen Quellen wie Giorgio Vasaris *The Lives of the Most Eminent Painters, Sculptors, and Architects* (1568). Solche Fälle verwischen die Grenze zwischen dem, was erkannt oder extrahiert wird, dem, was rekonstruiert wird, und dem, was erfunden wird.

Diese Dissertation entwickelt und evaluiert ein Framework, das den interpretativen Prozess als wesentliches Element der Wissensrepräsentation in die Extraktions-Pipeline integriert. Bestehende provenance-aware-Methoden erfassen in der Regel nur die Quelle, ohne die spezifischen Extraktionsbedingungen zu dokumentieren. Diese Lücke – die wir als Provenance Gap bezeichnen – ist grundlegend verschieden von Explainable-Artificial-Intelligence-Techniken, denn diese beantworten, wie ein Modell Eingaben verarbeitet, nicht aber unter welchen Bedingungen (Prompt-Templates, Modellversionen oder Reasoning-Schritten) eine konkrete Aussage produziert

wird. Was dort verloren geht, ist nicht eine technische Komponente. Eine Aussage, deren Entstehungsbedingungen nicht dokumentiert sind, kann weder rekonstruiert noch angefochten werden; die Verantwortung dafür geht im Graphen verloren.

Je mehr agentische Frameworks zum Standard werden, desto notwendiger wird eine umfassende Provenance-Dokumentation. Von Large Language Models erzeugte Reasoning Traces stellen zusätzliche Artefakte dar, die dasselbe Maß an Provenance-Dokumentation erfordern wie die Extraktionen, zu denen sie gehören. Diese Traces ergänzen den Provenance-Datensatz, ersetzen ihn jedoch nicht. Viewsari schließt den Provenance Gap, indem es Extraktionsbedingungen explizit verfolgt und eine klare Trennung zwischen textuellen Mentions und realen Referenten aufrechterhält, eine Unterscheidung, die Standards wie das CIDOC Conceptual Reference Model (CRM) häufig übergehen.

Als zentrale Fallstudie verwenden wir Giorgio Vasaris *The Lives*. Dieser bedeutende kunsthistorische Text verbindet historische Biografie mit narrativer Erzählweise; Kunstwerke werden selten unter ihrem offiziellen kanonischen Namen genannt, sofern sie überhaupt einen tragen, sondern vielmehr durch ihren Auftraggeber, ihren Standort, ihr Motiv oder ihren Zweck beschrieben. Da der Text implizite Referenzen verwendet, Werke enthält, die in kanonischen Datenbanken fehlen, und Beschreibung mit Interpretation vermischt, wo keine eindeutigen Fakten vorliegen, stellt er ein anspruchsvolles Testfeld für konventionelle Extraktions-Pipelines dar, das deren Leistungsgrenzen aufzeigt.

In dieser Dissertation werden vier Engineering-Artefakte konstruiert. Als Fundament dient die Viewsari-Ontologie, eine Provenance-aware OWL-2-Ontologie, die PROV-O und die W3C Web Annotation Ontology in einer dreischichtigen Struktur aus bibliografischer, struktureller und inhaltlicher Schicht integriert. Diese Architektur ist domänenneutral, mit Ausnahme der Inhaltsschicht, die auf die Kunstwerke des Fallbeispiels zugeschnitten ist. Die Ontologie legt drei zentrale Einschränkungen für die Pipeline fest: ein Zielschema, eine vierteilige Mention-Typologie (*explicit*, *implicit*, *coreferent*, *generic*), die Auflösungsstrategien trennt, und Provenance Slots, die Extraktionsschlussfolgerungen als PROV-O-Aktivitäten modellieren.

ObliquER (Oblique Entity Recognition), auf diesem Schema basierend, ist eine generative Extraktions-Pipeline mit einem modularen Prompt-System, das domänenspezifische Inhalte von strukturellen Templates trennt. Diese Modularität ermöglicht das systematische Testen und erleichtert die Anpassung an neue Domänen. ObliquER ist darauf ausgelegt, sowohl explizite als auch implizite Mentions zu erkennen und zu verknüpfen sowie Out-of-Knowledge-Base-Entitäten einzuschließen, statt sie zu verwerfen. Auf Entitäts-Ebene sind von den 15.804 Artwork-Entitäten im resultierenden Knowledge Graph 12.529 (79,3 %) Out-of-Knowledge-Base, d. h. sie lassen sich mit keiner Wikidata-Entität verknüpfen, während nur 3.275 (20,7 %) mit Wikidata verknüpft sind. Auf Mention-Ebene unterscheiden sich die beiden Typen deutlich: explizite, benannte Mentions

werden weitaus häufiger mit Wikidata verknüpft als implizite (49,0 % gegenüber 30,2 %), sodass implizite Mentions den Großteil der Out-of-Knowledge-Base-Population ausmachen.

Implizite Referenzen und deren Fehlen in zeitgenössischen Knowledge Bases korrelieren, denn die Entitäten, deren Erkennung Vasari bei seinen Lesenden voraussetzt, ohne sie explizit zu benennen, sind häufig genau jene, denen kanonische Identifikatoren fehlen. Eine Pipeline, die sich auf explizite Mentions beschränkt, würde daher nicht nur einen Teil der Artwork-Population ausschließen, sondern auch gezielt jene Entitäten ausschließen, die für die kanonischen Narrative konstitutiv sind, die *The Lives* selbst begründet hat.

ObliquER erfasst implizite Artwork-Mentions, die Surface-Form-Level-Zero-Shot-Baselines wie Universal NER und GLiNER2 aufgrund ihrer Architektur überhaupt nicht repräsentieren können. Auf dem impliziten Subset erzielen diese Baselines einen Strict F1 von 0,00 bzw. 0,02, gegenüber ObliquERs 0,321. Das Ergebnis ist daher eine Erkennungsfähigkeit für interpretative Quellen, nicht eine Verbesserung gegenüber einem bisherigen Bestwert. Beim Out-of-Knowledge-Base-Linking erreicht ObliquER einen F1 von 0,721, was zeigt, dass Out-of-Knowledge-Base-Mentions zuverlässig als solche erkennbar sind – eine operationale Voraussetzung dafür, sie im resultierenden Knowledge Graph als First-Class-Entitäten und nicht als Extraktionsfehler zu behandeln.

ObliquER erzeugt über dieses Schema den Viewsari Knowledge Graph (2.455.976 Triples, darunter 15.804 Artwork-Entitäten mit vollständiger Provenance für jede maschinell extrahierte Annotation), der zusammen mit seiner Weboberfläche zwei Funktionen erfüllt. Als Digital-Humanities-Werkzeug macht der Graph 12.529 bislang nicht katalogisierte Kunstwerke aus *The Lives* – Entitäten, die in Wikidata fehlen und für konventionelle Extraktions-Pipelines unsichtbar sind – recherchierbar und zitierbar; die Oberfläche ermöglicht es kunsthistorisch arbeitenden Personen ohne technischen Hintergrund, die Provenance jedes Sachverhalts einzusehen und in eigenen Arbeiten anzufechten. Als methodisches Beispiel zeigt ein Vergleich mit CIDOC CRM, dass Viewsari gleichzeitig ein Repräsentationsprinzip für implizite Entitäten und Extraktions-Provenance einführt, wie der Graph zeigt.

Um die konzeptionelle Übertragbarkeit auf andere Daten zu prüfen, passt ein Transferability-Pilot die Viewsari-Ontologie und ObliquER an ein deutschsprachiges Korpus des 19. Jahrhunderts über die Illuminaten (Kloß' Annalen) an, mit FactGrid als angestrebte Verlinkung anstatt Wikidata. Der Pilot grenzt ab, welche Komponenten allgemeingültig sind – die dreischichtige Architektur, die Mention-Typen, Pipeline-Stufen, Prompt-Struktur und Linking-Strategie – und welche domänenspezifisch sind. Eine vergleichende Analyse von Digital-Humanities-Projekten zeigt, dass es zwar einen Fokus auf Provenance-Repräsentation gibt, aber nur Viewsari Computational Provenance mit mention/referent-Trennung modelliert, das hier abstrahiert und als wiederverwendbarer

Beitrag zum Ontologiedesign in den Digital Humanities vorgeschlagen wird (das Computational-Provenance-Pattern) und die maschinell erzeugte Dimension generativer Extraktion erfasst, die andere Projekte nur informell dokumentieren oder ganz auslassen.

Ein Katalog mit 70 Competency Questions, entwickelt in Zusammenarbeit mit Kunsthistoriker*innen, verbindet diese Anforderungen mit der Evaluation. Alle Fragen sind auf dem befüllten Graph beantwortbar, und verbleibende Lücken identifizieren präzise fehlende Daten statt unbeantwortbarer Anfragen. Dieser Prozess klärt die Deliverables des Projekts für jedes Zielpublikum: eine abfragbare, quellenverankerte Antwort auf kunsthistorische Fragen sowie eine provenance-aware-Repräsentation des Extraktionsprozesses als First-Class-Element im Knowledge Engineering.

Der zentrale methodische Beitrag dieser Dissertation lautet, dass die Konstruktion eines Knowledge Graphs aus interpretativen historischen Quellen die Modellierung der interpretativen Arbeit an jedem Entscheidungspunkt erfordert, nicht lediglich die Extraktion oder Erkennung von Fakten. Der Prozess muss die Bedingungen dokumentieren, unter denen jede Extraktion erfolgt, und zwar mit derselben Sorgfalt wie die Sachverhalte selbst. Diese Dokumentation ist unverzichtbar, denn ohne sie geht die Verbindung zwischen einer Aussage und ihrem interpretativen Akt verloren, und mit ihr die epistemische Verantwortung für diesen Sachverhalt. Ein Graph, der nicht angibt, unter welchen Bedingungen eine Aussage getroffen wurde, lässt diese Aussage nicht gänzlich gegengeprüfen. Eine solche Dokumentation ist keine Form von Explainability im technischen Sinne; sie legt nicht die internen Berechnungen des Modells offen, sondern konstituiert hermeneutische Verantwortlichkeit, indem sie die textuelle Grundlage, die interpretative Konvention und den computationalen Agenten festhält, durch den jeder Sachverhalt produziert wurde. Darin überträgt sie die historische Quellenkritik auf die maschinelle Extraktion. So wie die Historiografie eine Quelle idealerweise nicht für bare Münze nimmt, sondern fragt, unter welchen Bedingungen, von wem und zu welchem Zweck sie entstanden ist, unterzieht Viewsari jeden maschinell extrahierten Sachverhalt derselben Prüfung und behandelt jede Extraktion als zu kritisierende Quelle, nicht als zu akzeptierendes Faktum. Viewsari demonstriert, dass eine solche Modellierung in großem Maßstab machbar ist und Interpretation zum integralen Bestandteil der Wissensrepräsentation macht, statt sie beim Übergang vom Text zum Graph zu verlieren.

Abstract

This dissertation argues that machine extraction from interpretive sources is an act of source production itself and, therefore, must be subjected to the same level of source criticism. It develops this principle into a formal framework for constructing knowledge graphs, in which every claim that a generative model extracts is treated as a source to be criticized, rather than as an accepted fact.

The central scholarly contribution lies in this insight that the construction of knowledge graphs from interpretive historical sources requires the explicit representation of interpretations and their provenance as independent units of knowledge in order to ensure traceability, critical evaluation, and scholarly transparency.

When large language models extract knowledge from interpretive historical texts, the output often seems like the kind of conclusions that human readers would otherwise draw. For instance, they must determine whether "the minister of culture to the Medici court in Florence" refers to Giorgio Vasari or whether a "small devotional panel" denotes a particular artwork. Conventional knowledge graph construction pipelines presuppose that such mention-to-referent resolutions are resolved prior to representation. This dissertation challenges this pre-interpretive entity assumption.

We emphasize this as a defect, especially for interpretive sources such as Giorgio Vasari's *The Lives of the Most Eminent Painters, Sculptors, and Architects* (1568). Such cases blur the line between what is detected or extracted, what is reconstructed, and what is invented.

This dissertation develops and evaluates a framework that incorporates the interpretive process within the extraction pipeline as an essential element of knowledge representation. Existing provenance-aware methods typically record only the source text for each claim, omitting the specific conditions under which the claim is extracted. This omission, which we term the provenance gap, is fundamentally distinct from the opacity targeted by explainable artificial intelligence techniques, which answer how a model processes input rather than under which conditions – such as prompt templates, model versions, or reasoning steps – a specific claim is produced. What is lost in this gap is not detail but accountability. A claim whose conditions of production go unrecorded cannot be reconstructed or contested, and responsibility for them is lost within the graph. As

agentic frameworks become standard, the need for comprehensive provenance documentation intensifies. Reasoning traces generated by large language models constitute additional artifacts that require the same level of provenance documentation as the extractions they accompany. These traces supplement, but do not replace, the provenance record. Viewsari addresses the provenance gap by explicitly tracking extraction conditions and maintaining a clear distinction between textual mentions and real-world referents, a separation often overlooked by standards such as the CIDOC Conceptual Reference Model (CRM).

We use Vasari's *The Lives* as our main case study. This important art-historical text blends historical biography with narrative storytelling, and artworks are rarely named by their official, canonical names (if they have one) and are instead described by who commissioned them, where they are, what they show, or their purpose. Because the text uses implicit references, includes works not found in standard databases, and blends description with interpretation, where no clear facts are to be recovered, it is a challenging test bed for conventional extraction pipelines, bringing their performance to a limit.

The dissertation constructs four engineering artifacts in sequence. At its foundation is the Viewsari ontology, a provenance-aware OWL 2 ontology that integrates PROV-O and the W3C Web Annotation Ontology within a three-layer structure spanning the bibliographic, structural, and content layers. This architecture is domain-neutral except for the content layer, which is tailored to the artworks in the case study. The ontology enforces three principal constraints on the pipeline, namely a target schema, a four-part mention typology (explicit, implicit, coreferent, generic) that separates resolution strategies, and provenance slots that model extraction reasoning as PROV-O activities. Central to this contribution is that historical descriptions should not be linked directly to world objects, but rather through explicitly modeled acts of interpretation. Viewsari implements this principle by separating the mention, the interpretation, and the referent.

ObliquER (Oblique Entity Recognition), built on this schema, is a generative extraction pipeline that employs a modular prompt system to separate domain-specific content from structural templates. This modularity enables systematic testing and facilitates adaptation to new domains. We design ObliquER to recognize and link both explicit and implicit mentions and to include out-of-knowledge-base entities rather than discarding them. At the entity level, of the 15,804 artwork entities in the populated knowledge graph, 12,529 (79.3%) are out-of-knowledge-base, meaning they cannot be linked to any Wikidata entity, while only 3,275 (20.7%) are Wikidata-linked. At the mention level, the two types diverge sharply: explicit, named mentions resolve to Wikidata far more often than implicit ones (49.0% versus 30.2% linked in the populated graph), so implicit mentions account for the bulk of the out-of-knowledge-base population. We find that implicit references and absence in contemporary knowledge bases track each other. Entities that Vasari expects readers to recognize without explicit naming are often those lacking canonical identifiers.

Restricting a pipeline to explicit mentions would therefore exclude not only a portion of the artwork population but specifically those entities most integral to the canonical narratives that The Lives itself established. ObliquER recovers implicit artwork mentions that surface-form-level zero-shot baselines such as Universal NER and GLiNER2 cannot represent. On the implicit subset, these baselines score a strict F1 of 0.00 and 0.02, respectively, by architectural construction, against ObliquER's 0.321, so the result reflects recognition capability for interpretive sources rather than an improvement over the prior best. On out-of-knowledge-base linking, ObliquER reports an F1 of 0.721, demonstrating that out-of-knowledge-base mentions are reliably detectable as such, an operational requirement for treating them as first-class entities in the resulting knowledge graph rather than as extraction failures.

Running ObliquER over this schema yields the Viewsari knowledge graph (2,455,976 triples, including 15,804 artwork entities, with complete provenance on every machine-extracted annotation), which, together with its web interface, serves two roles. As a digital humanities tool, the graph makes 12,529 previously uncataloged artworks from *The Lives* – entities absent from Wikidata and therefore invisible to conventional extraction pipelines – searchable and citable, and the interface lets non-technical art historians inspect and challenge the provenance of any claim. As a methodological example, a comparison with CIDOC CRM shows that Viewsari simultaneously adds a representation principle for implicit entities and extraction provenance, as the graph demonstrates.

To assess the application to other data, a transferability pilot adapts the Viewsari ontology and ObliquER to a 19th-century German corpus on the Illuminati (Kloß' Annalen), with the ground truth linking to FactGrid instead of Wikidata. The pilot delineates which components are general – the three-layer architecture, mention types, pipeline stages, and prompt structure – and which are domain-specific. Comparative analysis across digital humanities projects reveals a shared emphasis on provenance, but only the Viewsari ontology implements structures of computational provenance and mention/referent separation, abstracted and proposed as a reusable contribution to digital humanities ontology design (the Computational Provenance pattern), capturing the machine-generated dimension of generative extraction that other projects document only informally or omit.

A catalog of seventy competency questions, developed in collaboration with art historians and knowledge engineers, connects requirements to evaluation. All questions are answerable on the populated graph, and any remaining gaps precisely identify missing data rather than reflect unasked queries. This process clarifies the project's deliverables for each audience, namely a queryable, source-anchored response to art-historical questions and a provenance-aware, first-class representation of the extraction process for knowledge engineering.

The central methodological claim of this dissertation is that constructing a knowledge graph from interpretive historical sources requires modeling the interpretive work performed at each decision point, not merely extracting or detecting facts. The process must document the conditions under which each extraction is produced, recording them with the same rigor as the claims themselves. This documentation is essential because without it, the connection between a claim and its interpretive act is lost, and with it the epistemic responsibility for the claim. A graph that cannot say under which conditions it asserted something cannot be held accountable for the assertion. Such documentation is not a form of explainability in the technical sense; it does not expose the model's internal computations, but constitutes hermeneutic accountability, recording the textual basis, the interpretive convention, and the computational agent by which each claim was made. In this, it extends the historian's source criticism to machine extraction. Just as historiography never takes a source at face value but asks under what conditions, by whom, and to what end it was produced, Viewsari subjects each machine-extracted claim to the same scrutiny, treating each extraction as a source to be criticized rather than a fact to be accepted. Viewsari demonstrates that such modeling is feasible at scale, making interpretation integral to knowledge representation instead of something lost when moving from text to graph.

Preface

This dissertation is the result of research conducted between 2023 and 2026 at the Institute for Applied Informatics and Formal Description Methods (AIFB) at the Karlsruhe Institute of Technology (KIT) and at FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, within the Information Service Engineering (ISE), now also known as the Knowledge-Driven Artificial Intelligence (KDAI) research group led by Prof. Dr. Harald Sack.

This work began with a question that seemed straightforward at the time, back when I started the doctoral position right after my Master's thesis. Prof. Sack and I were discussing possible project ideas, and he suggested continuing to work on Vasari's Renaissance biography collection, knowing well that I have a passion for art history and even took it as a minor during my Master's studies. We discussed the direction and agreed that it would be interesting to investigate how the layered content of Giorgio Vasari's *The Lives* could be made accessible as structured, machine-actionable knowledge. I had glorious ideas for visualizations, a fancy web interface, and ways to enable exploratory knowledge graph queries. However, first we needed *data* to add to such a knowledge graph. It became clear that the extraction will be a ginormous challenge – and this challenge should be documented as such. We quickly realized that a knowledge graph built from an interpretive historical source is a mirror of the source it is derived from. Speaking to the domain experts and Vasari connoisseurs, we knew that Vasari constructed his own version of the Italian Renaissance in his work, and we were on the way to reproducing stereotypes and misconceptions that have been retold over 400 years in a user interface. In nearly every one of the sessions that followed, the same objection came back: *yes – but how are we supposed to believe any of what the machine does?* I took the question as motivation – and the result, one that fosters source criticism, where each result has a trail of where it *came from* – is what this dissertation now revolves around. It eventually became the central thesis: that in the age of generative artificial intelligence, modeling *interpretation* is as important as modeling the knowledge that interpretation yields.

Very early on, I had this idea that the project should be called *Viewsari*; a portmanteau of *view* and *Vasari* (... and also, kind of one of my hated nicknames, Sari). It reflects the dual concern of transforming a non-neutral catalog full of personal storytelling that has, over several hundred years, shaped how we see Renaissance art, and why the Teenage Mutant Ninja Turtles are named after Michelangelo, Leonardo, Raphael, and Donatello, rather than Pontormo, Ghirlandaio, Battista,

and Brunelleschi, into something called a ‘knowledge’ graph – a collection of *facts*. No, we would see the Renaissance through Vasari’s eyes. The project’s name is a reminder that what we build is not a mirror of the source but a *view* of it, one that should be documented as such.

I would like to express my gratitude to the people who made this work possible.

This project received support, expertise, and encouragement from many wonderful people. Heartfelt thanks to Prof. Dr. Anna Schreurs-Morét, Paula Schulze, Olivia Schmidt-Thomee, Dr. Grischka Petri, Dr. Ulrike Blumenthal, Annika Leuschner, Brechje Bakker, Etienne Posthumus, Dr. Georg Schelbert, and Thorsten Wübbena for their invaluable contributions. Some in particular participated in the requirements engineering workshops and the iterative ontology design process, engaging with complex formal knowledge representations and difficult questions.

I would also like to thank Dr. Mareike König, who supervised my Master’s thesis and later volunteered as a mentor. She is the person who, in academia, has pushed me the most by always believing in my potential and letting me be the free bird I am.

I thank Prof. Dr. Clemens Beckstein and his team for the support and feedback in stressful times.

Finally, I thank my family and friends for their unwavering support throughout the years of this work. This dissertation is dedicated to them.

To my mother, my number one supporter since 1999, who has always been there and never once stopped believing in me.

To Eva, who shares my love of books and matcha, and who helpes me rescue Rohes Ei in Heidelberg.

To Alex (and Norman), whose stream played in the background through countless hours, and who were an unexpected crisis hotline through more than one difficult moment.

To Sarah, whom I met at the very end of my Bachelor’s, in the MiMoText project, where the two of us grew like butterflies out of our cocoons. It clicked on our very first walk together, when we both walked so fast that everyone else was left miles behind. We share a name, the same music (Arctic Monkeys), the same left hand, and the same goofiness; we would not stand a chance against a single wasp. She has been my anchor ever since, and the two of us have been mistaken for sisters more than once – my sister from another mom.

To Raya, my best friend, whom I met online, yet who has stood by my side for two years now, weathering the hardest times with me through countless rounds of Dead by Daylight. Carrying struggles of her own, she set them aside for me again and again, giving me the very shirt off her back to keep me warm.

And finally, to my partner, Cedric, my muse through the most stressful times, without whom I would never have made it through any of this. This was the hardest thing I have ever done, and there were more times than I can count when I wanted to throw it all away; he is the reason I never did. We met at ISWC 2024 in Baltimore, among all possible places, and found our way back to each other later, and from the very beginning, it felt the way Halsey's *So Good* made me feel when I first thought of him, like *all the stars are fated*. In Plato's *Symposium*, Aristophanes tells of people once whole and then split in two, each wandering through life in search of the half that would make them complete again; with him, the searching stopped. He is also the one who showed me *Project Hail Mary*, and what has stayed with me is the feeling at its heart, of being entirely alien to a world and still entirely at home in it. Ours is a love told in songs: the *seven* by Taylor Swift he listened to and thought of me, and now Provinz's *Walzer*, which says what I cannot:

> Du bringst mich zum Lachen und ich mag, wie du lachst
> Wir warten auf Regen und sind doch überrascht
> Wir reden über Träume, Träume, die wir lieben
> Die uns bleiben, uns verfolgen
> Und wir tanzen in leeren Räumen
> Ja, wir drehen uns im Purple Rain
> Es ist Viertel vor vier, irgendwo in der Stadt
> Und du zeigst mir den Dreivierteltakt
>
> Und immer wenn du lachst, lache ich auch
> Ich schreie deinen Namen, ich schreie ihn laut
> Du weißt, wenn du weinst, halt' ich's nicht aus
> Ich glaube, wenn du stirbst, dann sterbe ich auch

He is every film I never saw but dreamed of, every four-leaf clover and every eyelash I ever wished upon.

Karlsruhe, November 9, 2026

Sarah Rebecca Ondraszek

Disclosure of generative artificial intelligence use

In accordance with the ACM Policy on Authorship[1], the Recommendations for the use of artificial intelligence of the Helmholtz Association, and the Handreichung *Umgang mit generativer Künstlicher Intelligenz* of the KIT-Fakultät für Wirtschaftswissenschaften, this section discloses the use of generative artificial-intelligence tools in the preparation of this dissertation. The disclosure is provided to be transparent. Research design, methodological choices, empirical claims, analyses, and scientific contributions of this work are the author's own, and every artificial intelligence-assisted output was reviewed and verified by the author before inclusion. Systems that form part of the research method itself are documented in Chapter 7.

Tools used. The following tools were used to prepare this dissertation:

- **Anthropic's Claude** (`https://claude.com`), accessed both as the `Claude Code` command-line agent and as a chat client, over the course of 2025 and 2026;
- **Grammarly Pro** (`https://www.grammarly.com`), used as a browser plug-in and editor integration.

Tasks for which artificial intelligence tools were used. Claude was used as an assistant for the following production-related tasks:

- LaTeX formatting, layout adjustments, typography, and the generation of section and reference labels;
- Language editing, including phrasing suggestions, copy-editing, and the rewording of complex or long sentences; the author is not a native English speaker, and Claude served as a non-native-author writing aid in these respects;

[1] ACM Policy on Authorship, `https://www.acm.org/publications/policies/new-acm-policy-on-authorship`.

- Iterative copy-editing and condensation passes over already-drafted chapters, conducted as a back-and-forth with Claude Code, in which sections were shortened, redundant analysis was removed, and prose was tightened while preserving the author's argument and claims; these passes included cross-checking statements, with every resulting edit reviewed and approved by the author;
- Bibliography structuring and the highlighting of missing metadata for cited references (e.g. filling in DOIs);
- Mock-reviewer feedback on individual dissertation sections, in which Claude was prompted to read a draft section as an external reviewer would;
- Implementation support for the Viewsari companion web application (via Claude Code as a coding assistant);
- Structuring of the project's GitHub repositories (directory layout, naming conventions, file organization) and drafting of `README` files that summarize the contents of scripts, datasets, and code modules, synthesized from the corresponding dissertation passages so that the repository documentation remains consistent with the written description of each component;
- Configuration and operational management of the project's server infrastructure, including Docker Compose definitions and Git workflow scripts;
- Generation of statistical figures rendered with `matplotlib`, `seaborn`, and comparable plotting libraries (e.g. corpus-statistics histograms, evaluation-result bar charts, confusion matrices).

Grammarly Pro was used for spelling and grammar checking and for surface-level rephrasing of complex sentences during writing.

Copyright and confidentiality. The source material (public-domain editions of Vasari's *The Lives*) and the author's own draft text contain no third-party personal or confidential data.

Responsibility. The author takes full responsibility for the entire content of this dissertation, including any text, code, citations, and figures. Any errors of fact, interpretation, attribution, or reference that may remain are the author's alone.

Contents

Kurzfassung . . . . . **i**

Abstract . . . . . **v**

Preface . . . . . **ix**

Disclosure of generative artificial intelligence use . . . . . **xiii**

Acronyms and symbols . . . . . **xxi**

Online Resources . . . . . **xxv**

1 Introduction and motivation **1**
1.1 Problem statement 4
1.2 Motivation 13
1.3 Research questions 15
1.4 Contributions 17
1.5 Publications 22
1.6 Dissertation outline 24

2 A bird's perspective: Theoretical and methodological foundations **25**
2.1 The age of generative artificial intelligence 25
2.2 When interpretation is shared between humans and machines 29
2.3 Epistemology – When cultural authority constructs a canon 30
2.4 Provenance and digital hermeneutics 32
2.5 Statistical vs. epistemic uncertainty in text analysis 35
2.6 Situating the work: digital humanities and the methodological gap 35
2.6.1 A brief overview of the digital humanities 36
2.6.2 Mind the methodological gap 37
2.7 From unstructured to structured textual (historical) sources 37
2.8 Moving towards structured representations of historical sources 41
2.8.1 Information extraction for creating knowledge representation 41
2.8.2 Knowledge graphs and ontologies in the context of unstructured sources 42

2.9 The forms of entities in textual sources 42
2.10 Bringing the worlds together: text-to-knowledge-graph pipelines 45

3 The Semantic Web: historical development and foundations 47
3.1 The World Wide Web 47
3.2 Ontologies and ontology languages in theory 48
3.3 Basics of ontology engineering 50
3.4 Basic criteria for ontology evaluation 51
3.5 Core principles of querying 52
3.6 The foundations of knowledge graphs 52

4 Text to knowledge: Foundations of information and knowledge extraction 55
4.1 A brief overview of the evolution of language models 55
4.2 What are the steps of information extraction? 56
4.2.1 Preprocessing 57
4.2.2 Named entity recognition 57
4.2.3 Coreference resolution 58
4.2.4 Entity linking 58
4.2.5 Relation extraction 59
4.3 Open and closed information extraction 59
4.4 From rule-based systems to transformers: an overview 60
4.5 Large language models and generative extraction (for historical sources) 61
4.6 Beyond the closed-world assumption: long-tail and out-of-knowledge-base entities 62
4.7 Evaluation of extraction systems for domain-specific and historical sources 63

5 State of the art in generative knowledge extraction and knowledge engineering 65
5.1 Generative knowledge extraction 65
5.1.1 Overview of contemporary large language model families 65
5.1.2 State of the art of generative information extraction 67
5.1.3 Prompt engineering for generative information extraction 71
5.1.4 Going obliquer: Existing approaches for implicit entity recognition 76
5.1.5 Limitations of generative information extraction and traditional metrics for their evaluation 78
5.1.6 Explainable and interpretable artificial intelligence 85
5.1.7 Generative entity linking 86
5.1.8 Of the long tail and resolving out-of-knowledge-base entities 88
5.1.9 Best of all worlds? Project example: Computer-assisted semantic text modeling 89

5.2 Knowledge engineering . . . 91
5.2.1 Advancements in ontology engineering methodologies . . . 91
5.2.2 Ontologies in the digital humanities . . . 92
5.2.3 Examples of knowledge graphs in the digital humanities . . . 99
5.2.4 Prior computational work on Giorgio Vasari's *The Lives* . . . 109
5.2.5 Limits of current approaches translated to requirements in Viewsari . . . 111
5.2.6 Knowledge graph visualization and interfaces for historical data . . . 114

6 "As long as art lives never shall I accept that men are truly dead:" Giorgio Vasari's *The Lives* . . . 121
6.1 Advantages of a case study approach . . . 121
6.2 Requirements for a case study corpus . . . 124
6.3 Publication history and details about *The Lives* . . . 124
6.4 Data foundation: sources and pre-processing . . . 127
6.4.1 Corpus and dataset construction from the Gutenberg project . . . 127
6.4.2 First statistical overview of the Viewsari corpus . . . 128
6.4.3 Text extraction and preprocessing of HTML content . . . 130

7 Methodology . . . 133
7.1 Proposed approach . . . 133
7.2 Corpus overview and baseline analysis . . . 134
7.2.1 Understanding Vasari: Co-occurrence statistics and person coverage . . . 134
7.2.2 Distribution across volumes . . . 135
7.2.3 Preliminary comparison of explicit entity type coverage . . . 139
7.2.4 Prototypical knowledge graph . . . 139
7.3 The Viewsari ontology: design process and iterations . . . 141
7.3.1 Viewsari ontology design phase I . . . 143
7.3.2 Viewsari ontology design phase II . . . 157
7.4 The ObliquER Pipeline . . . 177
7.4.1 Overview, motivation, and challenges . . . 181
7.4.2 Relation to the Viewsari ontology . . . 183
7.4.3 Formal task definition of entity recognition and linking . . . 186
7.4.4 Annotation strategy and ground truth construction . . . 189
7.4.5 Annotation environment . . . 200
7.4.6 Annotation guidelines . . . 203
7.4.7 Ground truth annotation process example in INCEpTION . . . 207
7.4.8 Entity recognition with ObliquER – spotting the unnamed . . . 214
7.4.9 Entity linking with ObliquER – a sea of unknowns? . . . 227
7.4.10 Extending the pipeline to a contemporary art-historical corpus: ArtPedia . . . 236
7.5 Knowledge graph population . . . 240
7.5.1 Bibliographic information . . . 240

7.5.2 Persons . . . 241
7.5.3 Co-occurrences . . . 241
7.5.4 Ground-truth mentions and entities . . . 242
7.5.5 ObliquER output integration . . . 244
7.6 Knowledge graph deployment and evaluation infrastructure . . . 246
7.6.1 Containerized deployment with GraphDB . . . 246
7.6.2 Materializing rule-based inference for offline evaluation . . . 247
7.6.3 Competency-question evaluation . . . 248
7.6.4 Reasoner-level evaluation with HermiT . . . 248
7.6.5 Derived data layer for the web application . . . 249
7.7 Website and user interface . . . 250
7.7.1 Architecture and technology stack . . . 252
7.7.2 Biography paragraph viewer . . . 253
7.7.3 Knowledge-base browser . . . 255
7.7.4 Knowledge graph explorer . . . 255
7.7.5 SPARQL query interface . . . 257
7.7.6 ObliquER pipeline demonstrator . . . 258
7.7.7 Design principles . . . 260
7.7.8 Continuous updates . . . 261

8 Results . . . 263
8.1 The Viewsari ontology: final state and instantiation . . . 263
8.2 ObliquER: entity recognition and linking results . . . 264
8.2.1 Final prompt templates . . . 264
8.2.2 Ground truth and ObliquER extraction statistics for the Viewsari dataset . . . 269
8.2.3 Ground truth and extraction statistics: ArtPedia . . . 270
8.3 Knowledge graph population results . . . 273
8.3.1 Population metrics: catalog . . . 275
8.3.2 Instance counts by ontology class . . . 277
8.3.3 Activity and agent inventory . . . 280
8.3.4 External entity inventory and out-of-knowledge-base count . . . 281
8.3.5 Population history: Run 1 to Run 2 . . . 282
8.4 Website and user interface: final implementation . . . 283

9 Evaluation . . . 287
9.1 Ontology evaluation . . . 287
9.1.1 Evaluation design . . . 288
9.1.2 Logical consistency . . . 288
9.1.3 Competency question coverage . . . 289

9.2 Comparative modeling study: Viewsari vs CIDOC CRM 292
9.2.1 The unifying claim: textual mention versus referent 293
9.2.2 The combinatorial objection and its limits 296
9.2.3 Co-occurrence modeling and the difference of 'interpretations' 300
9.2.4 The CRMinf scope boundary 300
9.3 Qualitative features: epistemic adequacy 308
9.3.1 Evaluation method for epistemic adequacy 308
9.3.2 Observable bindings in the populated graph 309
9.3.3 Descriptive grounding in the populated graph 310
9.4 ObliquER: baseline comparison and ablation 310
9.4.1 Metrics and statistical methodology 311
9.4.2 Baseline definitions for recognition and linking 314
9.4.3 Comparison results: entity recognition 316
9.4.4 Comparison results: entity linking 321
9.4.5 Cross-model evaluation 328
9.4.6 Ablation study 331
9.4.7 ArtPedia. Transferability assessment 338
9.5 Knowledge graph: measuring the population results 339
9.5.1 Verbosity: Triples per mention 340
9.5.2 External linking coverage by entity type 341
9.5.3 Bibliographic reachability 341
9.5.4 Provenance transparency 342
9.5.5 Methodological claims supported by the knowledge graph population 343

10 Beyond Viewsari: a transferability study on another source 345
10.1 FactGrid and the Illuminati order of Frankfurt as a controlled transferability pilot 346
10.2 The Knigge case study: applying ObliquER to the Kloß' *Annalen* 346
10.2.1 Corpus and source selection 346
10.2.2 ObliquER recognition results on the Kloß corpus 351
10.2.3 From recognition to a knowledge graph 352
10.3 Transferable design patterns – Viewsari for other humanities knowledge graphs 355
10.3.1 From case studies to comparative patterns 355
10.3.2 The Computational Provenance pattern 358
10.3.3 The co-occurrence content pattern 361

11 Discussion 363
11.1 When interpretation becomes an epistemic act of large language models 364
11.2 RQ1: Representational commitments for interpretive accountability 366

11.3 RQ2: Large language models as interpretive agents for implicit and out-of-knowledge-base entities . . . 370
11.4 RQ3: Design patterns and methodological transferability . . . 376
11.5 Complementary forms of interpretation: CRMinf and Viewsari . . . 378
11.6 On agentic large language models and the future of Viewsari . . . 380
11.7 Building a knowledge graph off of an 'unreliable' system . . . 381
11.8 Can the digital humanities and the related challenges offer insights on best practices as solutions? . . . 383
11.9 Current limitations of the Viewsari framework . . . 385

12 Summary . . . 387
12.1 Revisiting the research questions . . . 388
12.2 Summary of contributions . . . 393
12.2.1 Modeling . . . 393
12.2.2 Interpretation . . . 394
12.2.3 The age of GenAI. . . . 394
12.3 Future work . . . 395
12.4 A panel without a name: closing reflections on interpretive accountability . . . 396

A Appendix . . . 399

List of Figures . . . 427

List of Tables . . . 443

List of Publications . . . 453

Bibliography . . . 455

Acronyms and symbols

Acronyms

AAT	Getty Art & Architecture Thesaurus
ABox	Assertional Box (instance-level assertions in a Description Logic ontology)
AEVS	Anchor–Extraction–Verification–Supplement (extraction framework)
AI	Artificial Intelligence
API	Application Programming Interface
ATR4CH	Authenticity Assessment for Cultural Heritage (ontology)
BERT	Bidirectional Encoder Representations from Transformers
BFO	Basic Formal Ontology
CASTEMO	Computer-Assisted Semantic Text Modelling (framework)
CH	Cultural Heritage
CIDOC CRM	CIDOC Conceptual Reference Model
CIKM	ACM International Conference on Information and Knowledge Management
COMP-PROV	Computational Provenance (ontology design pattern)
CoNLL	Conference on Computational Natural Language Learning
CoT	Chain-of-Thought (prompting)
CQ	Competency Question
CSS	Cascading Style Sheets
CSV	Comma-Separated Values

DH	Digital Humanities
DL	Description Logic
DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
EL	Entity Linking
EOKG	European Olfactory Knowledge Graph
ExtrART	Annotation reference for art-historical information extraction (Santini et al. 2022)
FaBiO	FRBR-aligned Bibliographic Ontology
FactGrid	Wikibase-based knowledge graph for historical data
FAIR	Findable, Accessible, Interoperable, Reusable (data principles)
FRBR	Functional Requirements for Bibliographic Records
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
GT	Ground Truth
HiCO	Historical Context Ontology
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
ICONCLASS	Iconographic Classification System
IE	Information Extraction
IFLA	International Federation of Library Associations and Institutions
INCEpTION	Semantic annotation platform
InTaVia	In/Tangible European Heritage (project and platform)
IRI	Internationalized Resource Identifier
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
KB	Knowledge Base

KE	Knowledge Extraction
KG	Knowledge Graph
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
MEETUPs	Musical Meetups (Polifonia ontology module)
MemO	Memory Ontology (NFDI4Memory)
ML	Machine Learning
MUC	Message Understanding Conference
NeOn	Networked Ontologies (engineering methodology)
NER	Named Entity Recognition
NFDI	Nationale Forschungsdateninfrastruktur (German National Research Data Infrastructure)
NFDI4Memory	NFDI consortium for the historically oriented humanities
NLP	Natural Language Processing
ODP	Ontology Design Pattern
OOKB	Out-of-Knowledge-Base (entity)
OWL	Web Ontology Language
PIEA	Pre-Interpretive Entity Assumption
PMI	Pointwise Mutual Information
PON	Polifonia Ontology Network
PROV-O	PROV Ontology
QID	Wikidata item identifier (Q-number)
RDF	Resource Description Framework
RDFS	RDF Schema
ReQ	Requirement (Phase I: ReQI; Phase II: ReQII)
RQ	Research Question

SPAR	Semantic Publishing and Referencing Ontologies
SPARQL	SPARQL Protocol and RDF Query Language
SWRL	Semantic Web Rule Language
T5	Text-to-Text Transfer Transformer
TBox	Terminological Box (class and property axioms in a Description Logic ontology)
TEI	Text Encoding Initiative
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
XAI	Explainable Artificial Intelligence
XD	eXtreme Design (ontology engineering methodology)
XML	Extensible Markup Language

Online resources

This section lists the resources referenced and used throughout the dissertation. Unless otherwise noted, all URLs and DOIs cited in this dissertation – both in the list below and in the footnotes of the main text – were last accessed on June 14, 2026.

Project repositories (ISE / FIZ Karlsruhe)

`viewsari` Primary project repository. Contains the Viewsari ontology, a sub-module of the ObliquER extraction pipeline, preprocessing scripts, competency questions, the knowledge graph population steps, the knowledge graph, and documentation.
`https://github.com/ISE-FIZKarlsruhe/viewsari`

`vasari_nlp` Repository accompanying earlier work on *The Lives*. Contains reference data files used in the co-occurrence and entity analyses, including the artworks list and motif list.
Works list: `https://github.com/ISE-FIZKarlsruhe/vasari_nlp/blob/main/data/works.csv`
Motifs list: `https://github.com/ISE-FIZKarlsruhe/vasari_nlp/blob/main/data/motifs.csv`

`ObliquER` ISE / FIZ Karlsruhe fork of the ObliquER repository. Contains the ground-truth annotation data (organized by volume), the complete prompt-template catalog (including all recognition, entity-linking, and ablation-variant templates with their companion content files), and the materials for the FactGrid transferability pilot.
Ground-truth annotation data: `https://github.com/ISE-FIZKarlsruhe/ObliquER/tree/main/data/viewsari/ground_truth`
Recognition and ablation prompt templates: `https://github.com/ISE-FIZKarlsruhe/ObliquER/tree/main/data/viewsari/prompts`
Entity-linking prompt templates: `https://github.com/ISE-FIZKarlsruhe/ObliquER/tree/main/data/entity_linking`
FactGrid transferability pilot (Kloß *Annalen* transcriptions, ground truth, adapted prompts, and

demonstration RDF serialization): `https://github.com/ISE-FIZKarlsruhe/ObliquER/tree/main/data/factgrid`

Alternative (in case of problems with the ISE repo): `https://github.com/sarahondraszek/ObliquER`

`vasari_network` Repository and companion web page for the co-occurrence network analysis of *The Lives*, including the Vasari–Wikidata artist linking data.
Repository: `https://github.com/ISE-FIZKarlsruhe/vasari_network`
Network visualization: `https://ise-fizkarlsruhe.github.io/vasari_network/`

Viewsari web application

Viewsari web interface The primary access point for the Viewsari knowledge graph.
`https://viewsari.ise.fiz-karlsruhe.de`

Viewsari knowledge base namespace The internal namespace used for Viewsari-specific entities in the RDF serialization.
`https://viewsari.ise.fiz-karlsruhe.de/kb/1.0#`

Annotation and natural language processing tools

INCEpTION Semantic annotation platform used for manual annotation of the ground truth corpus.
`https://inception-project.github.io/`

Label Studio Open-source data labeling platform, referenced as an alternative annotation environment.
`https://labelstud.io/`

Tesseract / OCR-D Fraktur models Tesseract OCR with the OCR-D Fraktur-trained model (`deu_latf+Fraktur`), used to digitize the Fraktur-set Kloß *Annalen* for the transferability pilot.
`https://tesseract-ocr.github.io/` and `https://ocr-d.de/en/models.html`

Ontology engineering and visualization tools

Protégé Ontology editor used to author the Viewsari ontology across both design phases.
https://protege.stanford.edu/

HermiT OWL 2 DL reasoner (version 1.4.3) used to check the consistency of the Viewsari ontology.
http://www.hermit-reasoner.com/

WIDOCO Ontology documentation generator used to produce the published Viewsari ontology documentation portal.
https://github.com/dgarijo/Widoco

Graffoo Graphical notation for OWL ontologies, used for the class-and-property diagrams of the Viewsari ontology.
https://essepuntato.it/graffoo/

D3.js, Vis.js, and PyVis JavaScript and Python libraries used to render the co-occurrence and knowledge-graph network visualizations.
D3.js: https://d3js.org/
Vis.js: https://visjs.org/
PyVis: https://pyvis.readthedocs.io/en/latest/

Miro Graphical framework for visualizations of ontology-related content.
https://miro.com/

Ontology design patterns Ontology Design Patterns, actual entries currently unavailable and must be consulted via the Web archive, see, for example, the Provenance content pattern, referenced during the Viewsari ontology design process.
https://web.archive.org/web/20250219135335/http://ontologydesignpatterns.org/wiki/Submissions:Provenance

Software and infrastructure

GraphDB RDF triple store hosting the Viewsari knowledge graph behind the public SPARQL endpoint.
https://graphdb.ontotext.com/

FastAPI and OpenAPI Python web framework and the API specification standard used to implement the Viewsari web application's service layer.

FastAPI: `https://fastapi.tiangolo.com/`
OpenAPI: `https://learn.openapis.org/specification/`

nginx Reverse proxy serving the Viewsari web application.
`https://nginx.org/`

Hydra Configuration framework used to compose the ObliquER pipeline run configurations.
`https://github.com/facebookresearch/hydra`

Zotero Reference management software used to organize the dissertation bibliography and export the BibTeX source.
`https://www.zotero.org/`

External knowledge bases

FactGrid Wikibase-based knowledge base for historical research, hosted by the Gotha Research Centre, used as the entity-linking target and partial silver standard for the Kloß *Annalen* transferability pilot (Chapter 10).
`https://database.factgrid.de/`

Wikidata General-purpose collaborative knowledge base, the primary entity-linking target for the Viewsari corpus.
`https://www.wikidata.org`

Wikidata Query Service Public SPARQL endpoint over Wikidata, used for entity lookups and validation during entity linking.
`https://query.wikidata.org/`

Primary text sources

Project Gutenberg – *The Lives of the Artists* (de Vere trans.) The ten-volume Gutenberg edition of Vasari's *The Lives* in the English translation by Gaston du C. de Vere, used as the textual basis for the Viewsari corpus.
Main page: `https://www.gutenberg.org/`

Internet Archive – *Lives of the Most Eminent Painters, Sculptors & Architects* (de Vere trans., Medici Society) Digitized scans of the 1912–1915 Medici Society edition of the de Vere translation, used for the side-by-side facsimile display in the Viewsari web interface.

Linked from the volume-overview page above, the individual volume scans are hosted on the Internet Archive (e.g. volume 1, Cimabue to Agnolo Gaddi).
Volume 1 example: `https://archive.org/details/livesofmostemine01vasauoft`
Volume overview: `https://onlinebooks.library.upenn.edu/webbin/metabook?id=livespainters`

Bayerische Staatsbibliothek – Kloß, *Annalen der Loge zur Einigkeit* (1842, digitized) Digitized copy of Georg Kloß's *Annalen der Loge zur Einigkeit . . . zu Frankfurt am Main: 1742–1811* (Frankfurt am Main: Horstmann, 1842; shelfmark BSB10446364), digitized via the Münchener Digitalisierungszentrum, used as the primary source for the FactGrid transferability pilot (Chapter 10).
`https://www.digitale-sammlungen.de/de/details/bsb10446364`

Wikipedia dump (20 January 2019) Wikipedia XML dump used to construct the negative-example corpus for ObliquER training.
`https://archive.org/download/enwiki-20190120/enwiki-20190120-pages-articles.xml.bz2`

Web Gallery of Art (WGA) Reference database of Western European fine arts, used for artwork metadata cross-referencing in the entity linking component.
`https://www.wga.hu/index.html`

1 Introduction and motivation

Cultural heritage (CH) anchors shared meaning, and information technology has reshaped the way it is studied for decades, ranging from the internet and mass digitization to online tools (Saalmann 2020). The latest such shift is generative artificial intelligence (GenAI), namely large language models (LLMs), which can access, process, and produce content. However, while they seem brilliant at 'understanding' the content presented to them, they lack certainty and factual accuracy, especially in low-resource domains such as CH (Schimmenti et al. 2026, Graciotti et al. 2025). This is also reflected in the fact that the availability of digital versions of the data does not imply actual understanding of the content. So, CH materials may turn digital but remain unstructured; human-readable content is preserved, but systematic analysis is not yet possible (Stewart and Sinha 2025).

For example, for literary analysis, the requirement necessary for understanding texts is to access larger corpora of texts, with the possibility of moving between flexible modes of analysis, so-called 'scalable reading', that iteratively shift between patterns and detailed textual interpretation, allowing for an examination of broader structures and their local manifestations. This requires a shift in research from accessing individual documents to analyzing relationships across collections; a representation form that is both human-readable, interpretable, and machine-actionable. Additional modeling steps are required to transform digitized sources into *machine-actionable* knowledge. Machine-actionable, as a term, refers to the capacity of digital objects to be interpreted and processed by a computational agent without human intervention (Wilkinson et al. 2016). This enables tasks such as object type identification and determining its relevance to a search query with little or no human intervention (Miksa et al. 2021). For these tasks to be executable, it has become necessary to move from *digitization* to *digitalization*, the digital transformation (Rehbein 2017, Dobson 2026).

As such, knowledge graphs (KGs), graph-structured representations of entities and their relationships, have emerged as a central technology for structuring heterogeneous information and connecting dispersed collections. Knowledge extraction (KE), encompassing the recognition and classification of entities, that is, discrete things of interest such as persons, locations, or organizations, their linking to external references, and the extraction of relationships between them, is the puzzle piece completing the picture, since it provides the pipelines for transforming unstructured

text into machine-understandable, actionable, and interoperable representations. This enables access to hidden structures in historical sources at a new level, since individual mentions scattered across documents can be aggregated, cross-referenced, and systematically queried, thereby enabling forms of analysis that precisely meet the needs we have formulated above. However, most such approaches assume that knowledge can be extracted from text as a set of stable entities and facts, an assumption that holds in domains where information is expressed explicitly and terminology is standardized, but does not hold for large parts of the humanities (Stewart and Sinha 2025, Dobson 2026).

The title of this dissertation, *Modeling interpretation in the age of GenAI*, verbalizes this challenge. We do not claim to model what an LLM thinks when it processes a historical text. It claims something more tractable and, we argue, more scientifically necessary, namely to model the conditions under which an interpretation was produced, which model, which prompt, which text span, which agent configuration, at which point in time. Interpretation in GenAI is not directly observable, but the epistemic circumstances that produced it are. Without documenting these circumstances, extracted knowledge cannot be evaluated, reproduced, or held accountable. We argue that any extraction pipeline for these sources should treat its *interpretive* steps as primary elements to be documented, rather than as invisible processes that produce seemingly neutral facts. Put differently, extraction without documentation is epistemology without accountability.

The artifacts we develop in response are part of the *Viewsari* project, comprising a provenance-aware ontology for representing extracted knowledge from interpretive sources, an LLM-based extraction pipeline (*ObliquER*) for populating it, and the Viewsari KG, which instantiates the ontology on top of it. The remainder of this introduction situates these components within the conceptual framework that motivates them.

We develop the associated argument through an application to Giorgio Vasari's *The Lives of the Most Eminent Painters, Sculptors, and Architects* (1568) (Vasari 1568).[1] It is a, if not *the*, foundational text of art history as a discipline, comprising biographies of prominent Italian Renaissance artists such as Michelangelo Buonarroti and Leonardo da Vinci. Despite its age, the work continues to have an indisputable impact, as can be seen, for example, in the very recent series of online courses offered by the National Gallery in London.[2]

[1] Original title: *Le vite de' più eccellenti pittori, scultori e architettori*. Vasari originally published the first version in 1550 and later expanded and revised it, resulting in the 1568 edition, which includes additional biographies. We work with Gaston du C. de Vere's 1912 English translation (Vasari 2008) rather than the 1568 Italian original because the digital edition available through Project Gutenberg provides a stable, paragraph-level structure suitable for automatic processing, and because contemporary LLMs exhibit better-documented performance on modern English than on Early Modern Italian (Santini et al. 2025).

[2] Event: `https://web.archive.org/web/20260319123132/https://www.nationalgallery.org.uk/events/vasari-the-lives-and-afterlives-of-renaissance-artists-april-2026`

We note on terminology that the term *knowledge graph* is most commonly associated with structured, fact-oriented representations such as Wikidata (Vrandečić and Krötzsch 2014) or DBpedia (Auer et al. 2007), in which each triple asserts a claim about the world and the graph's value to a querying user lies in the assumed correctness of those assertions. Our approach does not rest on this assumption; we treat that departure from fact-orientation as a symptom of the problem this dissertation addresses rather than as a defect of the resulting graph. The resulting KG is, in a different sense, a structured, queryable record of *interpretive* statements on the Italian Renaissance by Giorgio Vasari expressed in *The Lives*, in which the unit of representation is not the fact but the attributed claim together with the conditions of its production. Every annotation in the graph is linked to the textual span from which it was derived, the activity that produced it, the prompt template under which the activity ran, the model version that generated the output, and the typological status of the mention itself (explicit, implicit).

We explicitly name this unspoken rule of conventional KG construction the *pre-interpretive entity assumption* (PIEA). It entails that mention-to-referent resolution can be treated as completed prior to representation, thereby populating the graph with established facts. Concretely, under this assumption, the node a pipeline extracts for the sculpture "Pietà" on a page of Michelangelo's biography is modeled as the sculpture itself – a real-world referent – rather than as a *mention* of that sculpture in a particular textual span, from which the referent must still be inferred. For interpretive sources, this assumption fails. Viewsari does not expect users to take the extracted content as fact; instead, it provides the apparatus to ask, for any claim in the graph, how it came to be and to question the result. This also implicitly resolves whether KE on such a highly interpretive text can be trusted. At the scale of *The Lives*, where no one can verify each extraction by hand, Viewsari makes visible not whether an output is correct but where it came from, relocating reliability to the source, where the practice of source criticism is equipped to weigh it.

This reframing is an inherent design choice from the dissertation's central argument. Knowledge extracted from interpretive sources is not discovered but constructed; therefore, the unit of representation cannot be the fact alone but must be the claim plus the documented act of its construction. ObliquER does not produce facts and is not designed to do so. Its focus is on the interpretive extraction process; its accuracy does not compare to that of a fact-retrieval system, but the provenance layer ensures that each individual assertion remains transparent at the level of its production.

1.1 Problem statement

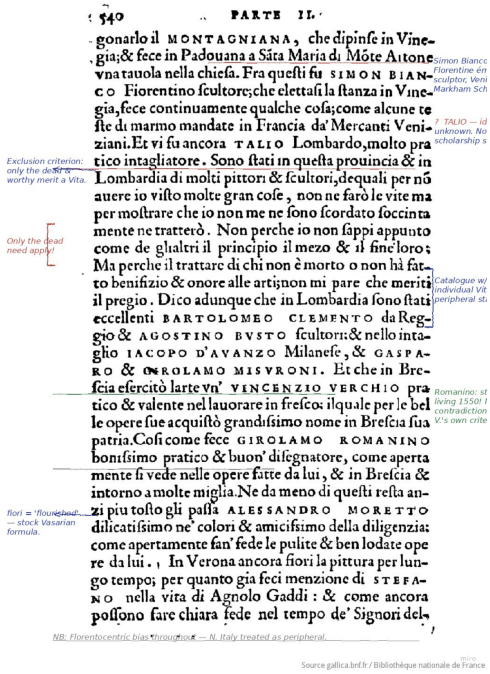
Imagine the following fictional scenario: an art historian frequents the archives of the Munich Central Collecting Point. The collection description features a portrait of a woman, possibly by Botticelli, with a note suggesting it was created between 1461 and 1510. When looking it up with an online image search tool, there is no title, and Wikidata entries thus define no label (see `Q111639719`). Surrounding context in the exhibition text suggests a connection to the Medici family, though the commissioner remains unnamed. The art historian knows there is a dispute regarding the provenance of this work; some believe it is indeed by Botticelli, others tentatively attribute it to his workshop, and there is no consensus on the exact date. For a computational system trained to extract data, what should be represented here?

Following up on this question, we establish the theoretical baseline for the dissertation in this problem statement, including a brief classification of the special nature of texts in the digital humanities (DH) and art history, an explanation of implicit entities and information extraction (IE), as well as an explanation of why LLMs bring with them a new level of problems.

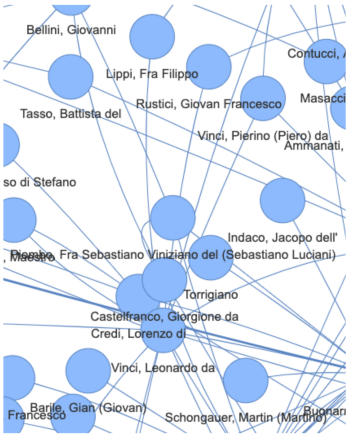
In historical (and thus art historical) textual sources, information is rarely presented as a quantifiable fact. Instead, relevant entities are embedded in descriptive language, and canonical names are often absent. Statements in such texts express attributions rather than objective observations. In traditional analyses of these documents, researchers draw on multiple sources to contextualize their interpretation of the content. This analysis, encompassing source criticism and contextualization, entails a targeted interpretation of individual passages and emphasizes interpretive insight (Ginzburg et al. 2013, Jänicke et al. 2015).

Digital methods, like the KE tasks we briefly mentioned before, such as named entity recognition (NER; identifying and classifying entity mentions in text) and entity linking (EL; connecting those mentions to knowledge base entries), have become a means to that end. Instead of doing just a pattern analysis of multiple text corpora at scale, as visualized in Figure 1.1, the goal is to extract units from the sources that can be further analyzed and recorded (Santini et al. 2022, Ehrmann et al. 2023, Drucker 2011).

This is precisely what Figure 1.2 illustrates. Close reading alone cannot scale across an entire corpus; distant reading alone cannot preserve context. A network of extracted names and co-occurrences, without interpretive qualification, produces a graph that is queryable at scale but epistemically opaque. Person A and Person B are connected, but whether that connection reflects a subjective attribution or a documented commission is lost entirely, as is the source context in which the relationship was recorded. KE occupies the necessary middle ground, transforming unstructured historical text into structured, machine-actionable representations with entity types,



(a) *Close reading*: a (fictional) historian's annotations on a page of a historical book (`https://gallica.bnf.fr/ark:/12148/bpt6k123255q/f541.item`), marking different entities.



(b) *Distant reading*: person co-occurrence network derived from the same book; nodes encode persons and edges represent shared contexts across the corpus.

Figure 1.1: Two complementary interpretive modes applied to the same historical text. Close reading (left) foregrounds the interpretive acts embedded in a single passage, and distant reading (right) reveals macro-structural patterns of co-mention across the full text that no single reader could perceive unaided.

qualified relations, and the provenance of every claim. KE brings the semantic layer to scalable reading.

However, there is not yet a perfect fit for such approaches to humanities texts; existing methods often treat texts as evidence of facts, whereas humanities research requires insights into resources as evidence of reasoning (Graciotti et al. 2025, Stewart and Sinha 2025). Consequently, the problem is not merely insufficient accuracy, as when a method fails to recognize the Botticelli painting we referred to above as a painting at all, but a fundamental mismatch between epistemic assumptions: factual knowledge representation versus interpretive knowledge construction.

Moreover, extracting knowledge from historical sources also requires recognizing entities that are never explicitly named but only described, the so-called implicit entities (Cavazza and Zweigenbaum 1992, Perera 2016, Stramiglio et al. 2025). Human communication is shaped as much by what is not said as by what is explicitly articulated. Implicitness, meaning that is suggested or

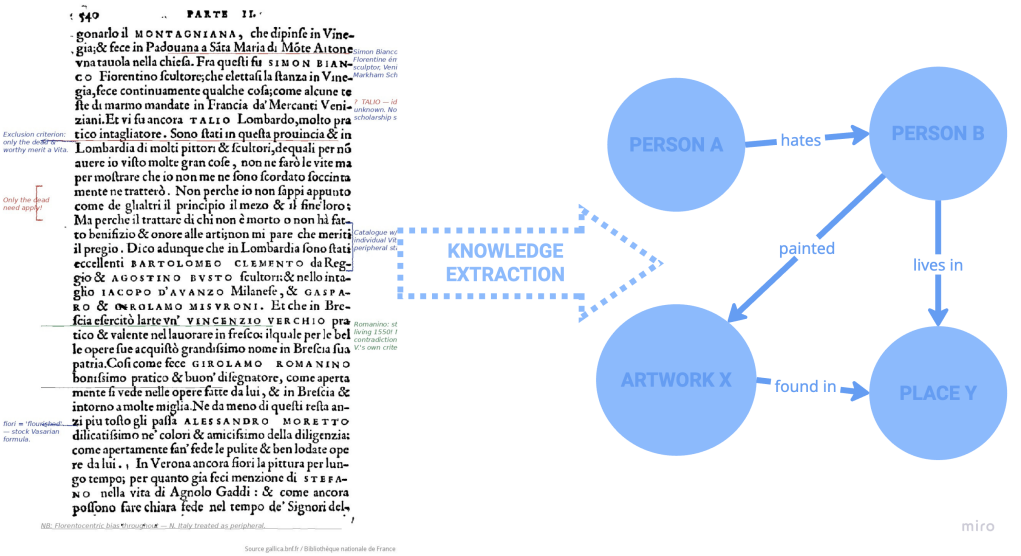


Figure 1.2: Close reading of a text yields annotations, such as persons, dates, and locations, which are often limited to small-scale spaces. KE, however, infers content and adds semantic meaning. In comparison to the network in Figure 1.1, the graph here interrelates persons based on semantically meaningful connections, such as 'found in'.

presupposed without explicit expression, affects different aspects of knowledge production. This influences how cultural knowledge is documented and preserved, not limited to spoken discourse alone, but manifesting across all forms of media throughout human history. These cues, which are often only understandable through shared context, allow an *oblique* transmission of meaning. However, implicit entities in the context of KE are potentially relevant entities that are missed by traditional methods because they lack a surface form and are only implied in the text (Hosseini 2019).

In the context of CH texts, implicit entities are particularly significant because their identification depends on contextual, historical, and often interpretive knowledge. A phrase such as "a small devotional panel once kept in the chapel" may denote a specific artwork, yet this identification arises only through inferential reasoning. Standard NER and EL pipelines, which rely on surface forms and candidate matching against existing knowledge bases (KBs), systematically overlook such references, especially in long-tail and out-of-knowledge-base (OOKB) scenarios (Hamaguchi et al. 2018). If an extraction method reduces these to simple name-relation-name triples, it does more than just lose detail. When these references are not resolved, it distorts how knowledge is represented.

The entities most often referenced implicitly are usually the most culturally important and, therefore, the most significant for understanding history, yet they may still be omitted from a KB. This is why, in traditional historical-critical editions, these mentions often come with footnotes that provide readers with additional context (Nova 2004). Early work on implicit IE already emphasized that recovering this type of content requires constructing a model of the described situation rather than relying solely on lexical triggers (Cavazza and Zweigenbaum 1992).

This challenge is compounded by the long-tail nature of CH data. Lesser-known artists and vague artwork references are precisely the entities that matter most for humanities research, yet they are systematically underrepresented in model training data and absent from standard KBs, rendering them OOKB entities (Graciotti et al. 2025, Hamaguchi et al. 2018).

In particular, in sources with linguistic variation and ambiguity, this remains a significant hurdle, and traditional natural language processing (NLP) techniques often fail in these domains. However, recent innovations in LLMs promise to bridge this gap (Stramiglio et al. 2025, Schimmenti et al. 2025, Santini et al. 2026). Importantly, surfacing these long-tail and OOKB entities enables scholars to recover knowledge that is absent from canonical or well-documented figures.

One of the goals of this dissertation is to recover this implicit semantic layer and make it representable within a KG. Rather than treating inferred entities as equivalent to explicitly named ones, we argue that their epistemic status must be modeled as interpretive assertions in the form of extracted mentions modeled through activities and resolved to referents correspondingly. Extracted entities are not discovered facts but results of a situated interpretive process based on textual mentions, be it performed by a human annotator or generated by a prompt-driven LLM (Schimmenti et al. 2026).

A core argument of this doctoral dissertation is therefore that when knowledge is extracted from humanities texts, several aspects occur simultaneously: statements originate from documentary sources, are produced through interpretive acts, and are generated by computational processes (Cavazza and Zweigenbaum 1992, Schimmenti et al. 2025). This is particularly consequential when extraction is performed generatively, that is, when an LLM – that is, a transformer-based autoregressive model trained on web-scale corpora via next-token predictions – produces entity descriptions and relations in response to interpretive instructions, rather than detecting surface forms already present in the text in a more traditional KE manner (Vaswani et al. 2017, Brown et al. 2020). In such settings, extracted statements are not detected directly in the text; rather, they are generated in response to instructions that specify the criteria for identifying an entity, relation, or reference (Zhang et al. 2025d). To clarify: in methods such as co-occurrence calculations using statistical measures like pointwise mutual information (PMI), knowledge is observed as statistical patterns already present in the text (Kalgotra and Sharda 2023), whereas in prompt-driven LLM extractions, knowledge is generated as the model synthesizes output in response to

interpretive instructions. Consequently, the resulting knowledge cannot be understood as a set of automatically observed facts; instead, it is shaped by the model's and its prompt's agency and is better characterized as interpretations produced by a strategized process (Xia et al. 2025).

Recent reflexive scholarship in NLP and computer science shows that prompt and model variability contribute to epistemic uncertainty, which means that small differences in prompt phrasing or design systematically influence the knowledge extracted and its perceived factuality, further establishing the interpretive rather than purely factual nature of KE in humanities applications (Xia et al. 2025, Dietrich and Hollstein 2025, Zhang et al. 2026). Both human annotators and computational systems thus reconstruct meaning rather than retrieve it, and variations in prompts or annotation guidelines produce systematically different results (Haq et al. 2025); prompt engineering accordingly functions not only as a technical procedure but as a form of interpretation, analogous to a historian's selection of methodological or hermeneutic frameworks (Phoenix and Taylor 2024).

This reframes extraction and prompt design from mere production tasks to epistemological processes. When interpretive assumptions are embedded in prompts and results without documentation, critical knowledge acquisition is undermined (Seddone 2025); provenance-aware extraction, by contrast, keeps the hermeneutic act visible, attributable, and contestable, since the extraction depends on prompt formulations, selection criteria, and contextual framing that operationalize interpretive assumptions about the source (Massari et al. 2026).

This requires explicit conceptual modeling. At this point, the need for a formal representation method becomes evident. Ontologies provide the necessary tools for this. An ontology, formally defined as an explicit specification of a conceptualization (Gruber 1993, Guarino 1995, Guarino et al. 2009, Hitzler et al. 2010), offers a formally defined vocabulary that specifies entities and relations, as well as the types of acts and provenance involved in the construction (Charles 2025). Without formalization, there is a risk of flattening the statements, rendering their origins and underlying reasoning processes untraceable. Ontological modeling makes this explicit, machine-interpretable, and interoperable (Gendron et al. 2026, Khorshidi et al. 2025).

This challenge is increasingly recognized: dedicated work in artificial intelligence (AI) epistemology and explainable artificial intelligence (XAI) examines the conditions under which outputs of generative systems can be considered reliable, attributable, and epistemically valid (Cambria et al. 2024, Bilal et al. 2025, Hohensinner et al. 2026). Earlier machine learning research approached accountability by interpreting model decisions – identifying which input features influenced a given prediction (Adadi and Berrada 2018, Arrieta et al. 2020). Generative models, however, have fundamentally shifted this problem: their outputs are not just selections from a predefined label space but constructions shaped by often oblique training corpora, by architectural properties such as scale and parameterization, and, crucially, by the instructions given at

inference time (Kaplan et al. 2020). More recent work in mechanistic interpretability attempts to address this by reverse-engineering LLMs' internal representations (Meng et al. 2022, Luo and Specia 2024). While this line of research has proven effective, it nevertheless addresses how a model processes information internally, not how interpretive instructions produce a specific output in a specific run. Documentation of prompts, model version, and interpretive criteria at the time of generation is indispensable; post hoc feature attribution and mechanistic analysis cannot fulfill this role. As extraction pipelines become increasingly agentic, meaning they also become tool-augmented, this requirement actually intensifies. A reasoning trace or an agentic lookup trail produced by an LLM constitutes a generated artifact, subject to the same stochastic variability as the extraction it seeks to justify (Plaat et al. 2025, Zhang et al. 2025a, Liu and Shu 2026). Recent grounded-extraction systems such as the Anchor–Extraction–Verification–Supplement framework operationalize precisely this insight by staging extraction across anchor discovery, anchor-constrained triplet generation, and restoration-based verification, precisely, and by reporting per-dataset hallucination rates as a direct measure of grounding success (Yang et al. 2026). This kind of grounded extraction, and the outputs taken as grounding evidence, must therefore be treated as an additional element within the provenance record, not as a replacement for primary documentation. This dissertation addresses this methodological imperative from the standpoint of historical scholarship, in which the absence of documentation directly compromises the credibility of claims about resources (Paulheim 2016, Schimmenti et al. 2026).

Unsurprisingly, the lack of transparency in interpretive claims has real-world consequences, and therefore, the structural necessity of provenance documentation has been thoroughly established in domains like the empirical sciences, where traceable provenance chains are a foundational requirement for the reproducibility of results and for maintaining trust in data pipelines (Gierend et al. 2024, Davidson and Freire 2008). These challenges are visible, for example, in surveys that examined cases of citations of retracted papers as legitimate, even long after retraction (West and Bergstrom 2021), and the spreading of misinformation of unanchored claims, such as the misappropriation of a five-sentence clinical letter into a decade-long justification for the opioid crisis (West and Bergstrom 2021).

Giorgio Vasari's *The Lives* also contains multiple examples that one can add to the list; one that is almost too on-the-nose appears in the 1568 edition, where he described his colleague Giovanni Antonio Bazzi as morally deviant, relying on oral reports and gossip rather than evidence. For over four centuries, most art historians accepted these views as fact and judged Bazzi's artistic work based on his supposed immorality (Dorey 2015).[3] A second, structurally similar case appears in the biography of the engraver Raimondi, in which he recounts a suit brought against

[3] German podcast episode on Bazzi: `https://www.kulturwissenschaften.uni-hamburg.de/ks/forschung/podcast.html`.

him by Albrecht Dürer. The anecdote has been transmitted as the first documented copyright case in European art history and continues to be reproduced – but the episode never took place. Petri (2014) demonstrates, on the basis of archival evidence and the chronology of Raimondi's copies, that Vasari assembled the story from disjoint elements, producing a *Phantomprozess*: a phantom case, narratively coherent and historically false, that has nonetheless structured the reception of both artists for nearly half a millennium. A third case appears in the biography of Michelangelo, where Vasari recounts how the Pietà in St. Peter's Basilica became the only work Michelangelo ever signed. According to Vasari, Michelangelo overheard a group of Lombard visitors praising the sculpture and attributing it to another contemporary Milanese sculptor (Vasari 1568).[4] Vasari reports that Michelangelo, unwilling to see his work misattributed, returned to the chapel that night and inscribed his name across the statue.

Clark and Holloway point out that these aspects constitute part of a larger paradox in the DH. While the 'traditional' humanities have always valued careful documentation and the authentic transmission of sources, DH projects rarely curate metadata on data provenance. As a result, the origins of digital knowledge are often as poorly documented as Vasari's verdict of Bazzi (Clark and Holloway 2012).

However, this information is central to scholarship. Whether it is an attribution, an inferred link, or a relation identified by an LLM, if there is no record of its production, it can enter the scholarly record as fact. Once it is included in KGs or cited in later work, tracing it back becomes nearly impossible.

Ironically, the position we contest in this dissertation is itself an implicit one. No one writes a paper about triples produced by a generative extraction pipeline being taken as facts. Yet the assumption is built into the dominant evaluation regimes and modeling conventions of the field, and it is in those conventions, rather than in any explicit claim, that the position is defended in practice. Schema-driven closed IE (Josifoski et al. 2022) frames the task as recovering a comprehensive set of triples, scored against a gold KG via exact match. The evaluation paradigm presupposes that extracted triples have a truth value independent of the conditions under which they were produced; the prompt, the model, and the extraction run vanish from the record once the triples are scored. For prose like in Vasari's work, where the act of identifying a referent is itself an interpretive judgment, the same regime silently flattens what extraction has produced. The same assumption is built into the apparatus by which the field measures itself: ontology-conformance benchmarks operationalize a category of *hallucination* defined as deviation from the input sentence's factual content (Mihindukulasooriya et al. 2023), a category that requires

[4] The passage is preserved in de Vere's translation: "one of them asked one of the others who had done it, and he answered, 'Our Gobbo from Milan' " (Vasari 2008). The episode is widely repeated in the scholarly literature; for a critical assessment of its historicity, see Unger (2014).

a fact of the matter against which model outputs can be measured – and that is precisely what interpretive sources do not supply.

The CIDOC Conceptual Reference Model (CRM) (Doerr 2003), the de facto standard for CH knowledge representation, follows this regime. When applied to pipelines that use text to create KGs about CH, the textual occurrence and the real-world object often collapse into a single node, and the textual evidence disappears from the graph. Recent work in CH has begun to push against this default: Schimmenti et al. (2026) reify scholarly disagreement explicitly. Our claim is not that this body of work is wrong on its own terms. Schema-driven extraction succeeds where it was designed to succeed, on explicit, well-curated, factually grounded sources. CIDOC CRM has provided a vocabulary that has organized CH data for two decades. What we argue is more limited and more specific: the assumptions that make these tools coherent for explicit sources are precisely the assumptions that fail for interpretive ones, and the failure is not a matter of accuracy but of representational kind. The layer this dissertation addresses is one step beneath the existing interpretive-aware work: not the interpretive activity of the human scholar, which the cited projects already document, but the modeling of the interpretive activity of the extraction pipeline itself, captured not as an extension or annotation but as a first-class component of the schema. The Viewsari ontology is the representational response to this gap, and we briefly introduce it in the following figures (Figure 1.5) before formalizing it in Chapter 7.

From interpretive claims to artwork mentions. The Bazzi and Raimondi cases demonstrate that provenance challenges in humanities extraction extend across entity types. Be it characterizations of people or descriptions of artworks, different cases face the same issue when extraction systems treat statements as facts. To tackle this challenge, this dissertation focuses on a single type of reference, namely mentions of artworks in Vasari's *The Lives*. Artworks are a challenging test case because a substantial fraction of the works Vasari discusses predate the establishment of systematic titling conventions (Welchman 1997, Yeazell 2015). Consequently, many such works lack canonical titles and are referenced in art-historical literature through descriptive phrases, artist attributions, or by location and commissioner. Indeed, even contemporary cataloging standards in the CH domain acknowledge that a "title" is not necessarily a canonical,

fixed reference but may take the form of an extended descriptive phrase, an iconographic identification, or an implicit designation by object type, dedication, or location.[5] EL systems that depend on surface-form similarity to knowledge-base labels cannot reliably resolve such references, as the absence of standardized titles prevents direct mapping. Recovering an artwork mentioned from such a description is therefore not just about recognizing words, but about making contextual inferences. This task is still done by human scholars, but is increasingly handled by generative systems. The framework developed here is intended to transfer beyond artwork extraction; as the Bazzi and Raimondi cases illustrate, the same provenance-aware methodology applies to other forms of attribution. This broader application is discussed further in Chapter 10.



Figure 1.3: René Magritte, *La trahison des images* (1929). Oil on canvas, 60,33 × 81,12 cm. Los Angeles County Museum of Art. Reproduced as a visual quotation in accordance with § 51 UrhG. © VG Bild-Kunst, Bonn 2026.

The treachery of mentions. The focus on artwork mentions, as we established above, is not a matter of methodological convenience; rather, it exposes a structural assumption inherent in

[5] The Getty's *Categories for the Description of Works of Art*, `https://www.getty.edu/publications/categories-description-works-art/`. It is a standard reference in the CH domain. Its Object/Work-Type category establishes the classification of an item as an artwork in the first place, distinguishing categories such as painting, sculpture, or architectural element (`https://www.getty.edu/publications/categories-description-works-art/categories/object-architecture-group/2/`). Its Titles category, in turn, explicitly covers both titles proper – descriptive phrases referring to iconographic subject (e.g., *Adoration of the Magi*) – and names functioning as titles for works lacking one, such as object types (*Ceramic Bowl*) or building dedications (*Saint Paul's Cathedral*). See `https://www.getty.edu/publications/categories-description-works-art/categories/object-architecture-group/3/`.

the representation of knowledge in the DH. A famous example of this is Magritte's *La trahison des images* (1929), a painting of a pipe, accompanied by the inscription *Ceci n'est pas une pipe* – "This is not a pipe". With this painting, Magritte visualizes the relationship between a sign and the thing it depicts, since the image itself is a pipe, only a representation of one (Foucault 1983). The same non-identity holds, with sharper consequences, for textual mentions of artworks in historical sources.[6]

Assuming PIEA, when ontological commitment is existential, every instance presupposes that the entity it represents exists, or has existed, in the world. This works well outside of IE pipelines. There, however, the extraction process never 'sees' "the pipe" but merely its mention, which is surfacing in a document, written by someone, at some time, in some genre, with some rhetorical purpose. Treating that mention as if it directly instantiates an object in the real world collapses referent (the entity in the world, if it exists at all) and its mention (the textual occurrence). We aim to invert this ordering by asserting the textual mention unconditionally (a character span exists in the source text, trivially), and the link to a referent is a result of a downstream inference step that may or may not succeed. The referent is optional; the mention is not.

1.2 Motivation

Based on the problem statement presented, we argue in this dissertation that knowledge derived from historical texts should not be modeled as a collection of objective facts but rather as the product of interpretive processes. Humanities research necessitates a representation that incorporates the agents, both artificial and human, and variants of the source, such as different editions of books, involved in this extraction process. To address this, we propose a methodology that models extraction as an interpretive activity and records the conditions under which it occurs, including prompts and software agents, as part of the provenance of the resulting KG. This methodology encompasses ontology design conducted in collaboration with domain experts – whose recurring objection to any fact-oriented representation, that an untraceable claim about a contested source cannot be believed, is itself a statement of source criticism and the reason provenance must be constitutive rather than ancillary – and treats KE as an interpretive process rather than a mere search for factual units in a text.

[6] We use *referent* and *mention* in their standard linguistic and philosophy-of-language assumption, where the referent is the entity and the mention is a particular textual occurrence of a referring expression (Frege 1948), corresponding to usage in coreference resolution and EL (Ehrmann et al. 2023).

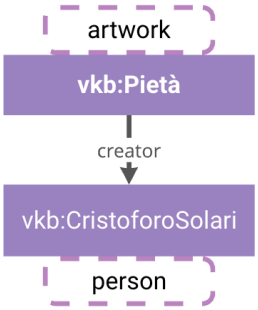


Figure 1.4: Naive KE applied to the Pietà authorship claim in Vasari's *The Lives*: the overheard attribution to Cristoforo Solari ("Our Gobbo from Milan") is encoded as a flat relation, with no record of its origin.

A motivating example. The example from before, about Vasari's Phantomprozess, makes evident what happens when claims can not be traced; Vasari's subjective narratives have influenced how art historians have perceived both artists for nearly half a millennium.

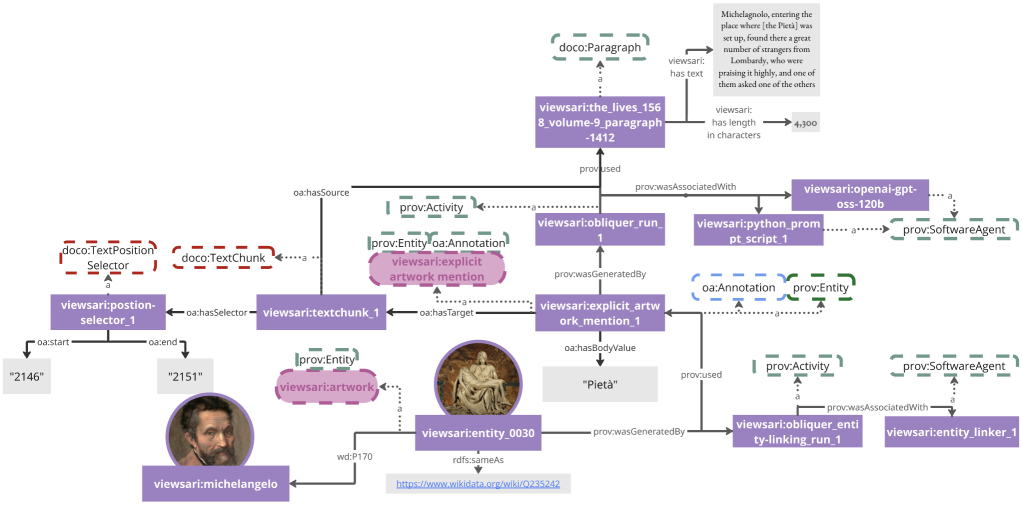


Figure 1.5: The same passage modeled in Viewsari's content layer: the explicit mention `Pietà` as a typed `oa:Annotation`, anchored to its source paragraph via `oa:hasTarget`, generated by a documented extraction run, and linked to `wd:Q235242` via `owl:sameAs`, where the actual creator (Michelangelo, `wd:Q5592`) is recorded. The contested Gobbo claim does not enter the KG as a creator triple.

In parallel, for Michelangelo's Pietà, a naive extraction system might encode the sentence as `[artwork: Pietà] [creator: Cristoforo Solari]`. But where is the record on this claim's origin, that it was an overheard remark reported by Vasari, who retells it fifty years later? Figure 1.4 shows the contested authorship claim entering the KG as a flat triple, structurally

indistinguishable from a documented commission. Figure 1.5 shows the same passage modeled in Viewsari, where the mention is typed as an explicit artwork mention, anchored to its source paragraph, generated by a specific extraction run, and linked via `owl:sameAs` to the Wikidata entry where the actual creator is documented.

This difference captures the dissertation's main methodological focus. The same source passage and the same extracted content yield two fundamentally different epistemic objects.

Thus, Vasari's work serves as a case study for the methodological claims of this dissertation because canonical titling conventions for Renaissance artworks did not emerge until after Vasari's time. As a result, Vasari implicitly refers to artworks, often describing their content, commission status, function, or location rather than using canonical titles. This lack of canonicity leads to systematically incomplete coverage of these entities in external KBs, where many canonical works described do have entries in Wikidata, but a significant proportion are either absent or listed under post-Vasarian titles that differ substantially from Vasari's descriptions in surface form. In such cases, the boundaries between recognition, identification, and uncertainty collapse into interpretive judgment. This convergence necessitates a provenance-aware approach to representation.

This choice also motivates the names of two of the four main contributions of this dissertation, which we elaborate on below. The name Viewsari is a pun on *view* and *Vasari*, emphasizing that using Vasari's work to construct a KG represents a *view* of the Renaissance art world through Vasari's lens, not an objective mirror of historical reality.

Knowledge extracted from interpretive texts is not discovered but constructed; therefore, knowledge graphs must model and make visible the construction process itself.

1.3 Research questions

Three core research questions (RQs) define the scope of this dissertation. Each RQ is answered jointly by an empirical artifact and one or more conceptual claims that are introduced in Section 1.4. These RQs are formulated in the tradition of design science research (Hevner et al. 2004, Wieringa 2014a) and are analogous to ontology engineering methodologies such as eXtreme Design (XD) (Presutti et al. 2009, Blomqvist et al. 2016), in which they articulate a representational or methodological gap and are answered through the construction of an artifact together with the design rationale supporting it. The wording of each question also adheres to the practical criteria of clarity, focus, specificity, researchability, relevance, and originality emphasized in standard guidance on constructing research questions.

RQ1: What representational commitments must a knowledge graph make to remain accountable to the interpretive act that produced it?

Thesis. We hypothesize that accountability to the interpretive act requires an explicit rejection of the PIEA (C1) and a three-layer architecture (bibliographic, structural, and content) that maintains the distinction between mention, referent, and bibliographic source. We hypothesize further that a provenance-aware ontology integrating PROV-O (Lebo et al. 2013) and the W3C Web Annotation Ontology (García Castro et al. 2010, Sanderson et al. 2017) can represent (a) explicit and implicit entity mentions, (b) the extraction activities that produced them, and (c) their grounding in specific textual passages across editions and translations, without flattening interpretive uncertainty into seemingly factual assertions. We hypothesize, finally, that these commitments are *not* natively expressible in the CH vocabulary most widely adopted in the domain and closest to Viewsari – the CIDOC CRM family – and therefore call for a dedicated provenance-aware ontology rather than an extension of an existing one. RQ1 is addressed conceptually by C1 and empirically by E1, with the question of whether existing vocabularies can satisfy these commitments tested through the comparative modeling study of Section 9.2.

Criterion of answerability. RQ1 is answered once the required commitments are named, shown to be unmet by the CIDOC CRM family (Section 9.2), and operationalized in an ontology with a schema that is logically consistent under OWL 2 DL reasoning, covered by the competency questions (CQs), and provenance-complete on the populated graph (Section 9.1). Full-graph reasoner classification is computationally out of scope (Section 9.1.2); the consistency requirement is therefore terminological, and graph-level answerability is established through CQ coverage rather than through description-logic entailment.

RQ2: Under what conditions can LLMs recognize and link the implicit and OOKB mentions that surface form-level baselines miss?

Thesis. We hypothesize that generative extraction introduces a distinctive representational deficit – the provenance gap (C2) – that any pipeline targeting interpretive sources must address as a design requirement rather than as a post hoc transparency layer. Furthermore, we claim that this addresses a conceptually distinct problem than XAI techniques and therefore demands a hermeneutic rather than mechanistic provenance (C3). We propose a modular prompt architecture that separates domain-specific content from structural templates, enabling LLMs to extract both explicit and implicit entity mentions and yielding measurable improvements over zero-shot / off-the-shelf NER baselines (such as UniversalNER (Zhou et al. 2024c)) on interpretive historical sources, particularly in recognizing OOKB and implicit mentions. LLMs are themselves treated

as interpretive agents that follow a documented strategy (Kasner et al. 2026).[7] RQ2 is addressed conceptually by C2 and C3 and empirically by E2 and E3.

Criterion of answerability. RQ2 is answered once a documented LLM pipeline recognizes and links implicit and OOKB mentions at a rate that surface form-level baselines cannot reach – and that gap holds even against baselines capable in principle of implicit recognition – with the gain attributable to the ontology-typed modeling rather than to incidental prompt tuning (Section 9.4).

RQ3: What reusable ontology design pattern emerges from modeling extraction as a documented, provenance-aware interpretive process, and *how* does that foundation transfer to structurally distinct (interpretive) corpora?

Thesis. We hypothesize that (a) a comparative ontology-pattern analysis across DH KGs reveals a novel pattern specific to generative extraction (C4), and (b) the Viewsari pipeline architecture can be adapted to structurally different source material – demonstrated through a pilot study on 19th-century German documentary sources – by adapting only the prompt content layer and the domain-specific mention subclasses (declared under existing parent classes), while holding the ontology core, three-layer architecture, templates, and post-processing fixed. RQ3 is addressed conceptually by C4 and empirically by E4.

Criterion of answerability. RQ3 is answered once the comparative analysis isolates a reusable design pattern unique to generative extraction and the pipeline is shown to transfer to a structurally distinct corpus by adapting only the prompt content layer and a small set of domain-specific mention subclasses declared under existing parent classes, with the ontology core, three-layer architecture, templates, and post-processing held fixed (Chapter 10).

1.4 Contributions

The central contribution of this dissertation can be stated concisely: it advances historical source criticism by applying it to machine extraction. It introduces an ontology that serves as the structural foundation for KG construction, treating each claim extracted by a generative model as a source subject to critical evaluation rather than as a fact to be accepted. The remainder of this section unfolds this into four conceptual and four empirical contributions.

Scope of the empirical contributions. This dissertation makes four conceptual and four empirical contributions. The conceptual framework, formulated based on recommendations

[7] LLMs as interpretive agents are not the same as agentic LLMs; see Shanahan et al. (2023) and Salimpour et al. (2025).

from Wieringa (2014a,b), articulates the representational and epistemological commitments that interpretive sources impose on a KG; the empirical contributions develop the artifacts that operationalize those commitments.

At this stage, the ObliquER pipeline, the KG, and the empirical evaluation are configured for artwork mentions in Vasari's *The Lives*, which constitutes the primary empirical case of this dissertation. This choice is a deliberate methodological decision, not a limitation of the framework. Artworks in pre-modern art-historical writing have structural features that make the recovery of implicit references particularly demanding and, therefore, constitute the configuration in which the central RQs can be evaluated most rigorously.

All Viewsari sources are accessible on GitHub at `https://github.com/ISE-FIZKarlsruhe/viewsari`.

Conceptual contributions.

C1: The pre-interpretive entity assumption (PIEA).
We articulate the PIEA as built into the dominant evaluation regimes and modeling conventions of IE, without ever being explicitly named. The PIEA states that mention-to-referent resolution can be treated as completed prior to representation. We show that PIEA fails systematically for interpretive sources, in which a substantial proportion of references are implicit, and we propose a three-layer ontology architecture (bibliographic, structural, content) that maintains the distinction between mention, referent, and bibliographic source as the representational response.

C2: The provenance gap in generative information extraction.
We identify a representational deficit specific to generative extraction – the *provenance gap* – and argue that provenance documentation is not a post-hoc transparency measure but a constitutive condition of scholarly assertion under generative extraction. This is a claim about *visibility*, not about kind: documenting the interpretive conditions under which a source is read has continuously been part of sound scholarship, but where classical pipelines left those conditions at least inspectable after the fact, generative extraction makes explicit, machine-recorded documentation the only remaining way to keep them recoverable. What was good practice therefore becomes, under generative extraction, a precondition of accountable assertion rather than an optional addition.

C3: Hermeneutic provenance addresses a conceptually distinct problem than XAI.
We argue that the provenance required by humanities scholarship is *hermeneutic*, not mechanistic, and that XAI techniques cannot substitute for it. Feature-attribution and mechanistic-interpretability methods answer which internal computations or input tokens

influenced this output; humanities readers ask on what textual basis, under which interpretive convention, and by which agent was this claim made. C3 is the methodological response to the gap identified in C2. In disciplinary terms, hermeneutic provenance extends historical source criticism – the methodical interrogation of a source's origin, agency, and conditions of production, codified by Bernheim (1908) and Droysen (1868) – by applying computational methods. The same critical apparatus that historiography brings to a human-authored source is here turned on the machine extraction itself, so that each extracted claim enters the KG as a source to be criticized rather than as a fact to be trusted; doing history in the digital age makes this turn unavoidable rather than optional (Weller 2013).

C4: Computational Provenance (COMP-PROV) as an ontology design pattern.
Through a comparative pattern analysis across four DH KGs, we identify a need for the abstraction of a recurring ontology design pattern (ODP) – Computational Provenance (COMP-PROV) – that reifies the extraction agent, prompt, model version, and run as first-class entities linked to every extracted assertion. We show that COMP-PROV is a genuine addition to the existing DH pattern repertoire for when LLM-based extraction populates a humanities KG; the comparative analysis itself constitutes a methodological contribution by establishing pattern-level requirements for generative extraction.

Empirical contributions.

E1: The Viewsari ontology.
An OWL 2 DL ontology that operationalizes C1–C3 as an annotation-and-provenance model grounded in textual mention spans and computational extraction processes. Viewsari integrates PROV-O (Lebo et al. 2013) and the W3C Web Annotation Ontology (Sanderson et al. 2017) along the three-layer architecture introduced in C1, representing extracted entities with the interpretive conditions of their production: extraction agents, prompt templates, and mention typologies (explicit, implicit, coreferent, generic). Rather than replacing established ontologies, Viewsari establishes a layered representation that ensures extracted knowledge remains traceable to both its textual source and the interpretive procedure that produced it. In the documentation stack it sits upstream of the CIDOC CRM family, supplying the recognition and entity-linking layer over text on whose resolved referents a CIDOC event layer can subsequently build (Section 11.5). It stands in dialogue with existing provenance and argumentation vocabularies in the CIDOC CRM family, most notably CRMinf (Doerr et al. 2023), which models propositional beliefs that actors hold about observable entities within an inference-making framework. Section 9.2 examines this relationship in detail and shows that CRMinf and Viewsari address different layers of interpretation and epistemic provenance, since CRMinf operates at the level of belief and

proposition, whereas Viewsari operates at the level of textual mention, extraction process, and annotation.

E2: The ObliquER pipeline.
A modular generative extraction pipeline for implicit and explicit entity recognition and linking that instantiates C2 at the system level: every extraction step is recorded as a PROV-O activity in the resulting graph. ObliquER separates domain-specific prompt content (entity definitions, mention typology, and few-shot examples) from structural templates, enabling systematic ablation and adaptation. The pipeline harmonizes entities linked to external KBs such as Wikidata, while assigning stable Viewsari-internal IRIs to OOKB entities and representing them as first-class KG nodes.

E3: The Viewsari KG and web interface.
A populated KG comprising 2,455,976 triples and 112,113 annotations under 100% provenance coverage, of which 86,654 are typed mention annotations – 84,215 produced by ObliquER and 2,439 from the manually annotated ground truth – deployed with a web interface that enables non-technical exploration of extraction provenance (Waitelonis and Sack 2012). The interface makes the abstract claim of C2 – that provenance must remain visible at the point of use – concretely inspectable, and constitutes the empirical demonstration that hermeneutic provenance (C3) is implementable at production scale.

E4: A transferability pilot on 19th-century German documentary sources.
The Viewsari ontology and ObliquER pipeline are reconfigured for person, location, and organization extraction from a 19th-century German-language documentary corpus on the Illuminati order, with FactGrid (Simons 2024) as the target linking knowledge base. The ontology core, three-layer architecture, templates, and post-processing remain unchanged; only the prompt content layer and the domain-specific mention subclasses, declared under their existing Viewsari parent classes, are adapted. End-to-end linking against FactGrid is reserved for future work; the pilot's contribution is at the level of recognition and representation. The pilot operationalizes C4 by demonstrating that COMP-PROV is not an artifact of the Vasari case but rather a transferable design commitment, thereby validating the recognition-and-representation axis of RQ3's methodological transferability claim.

Together, E1–E4 constitute the artifact-level realization of C1–C4: the ontology (E1) operationalizes PIEA's representational response and the hermeneutic-provenance commitment; the pipeline (E2) and graph-with-interface (E3) make provenance documentation concrete and inspectable; and the transferability pilot (E4) demonstrates that the resulting design commitments compose into a reusable methodology rather than a Vasari-specific artifact.

Architectural positioning. Architecturally, we position Viewsari in the neuro-symbolic space (Hitzler et al. 2024, Prenosil et al. 2025): the ontology (symbolic component) defines the epistemic baseline, thereby also formalizing and constraining the extraction tasks assigned to the LLM (neural component) (Khorshidi et al. 2025, Luong Tuan 2026). To make sources representable in a KG, the transformation pipeline includes tuning LLM prompts to recognize implicit entities in descriptions and evaluating potential models for entity matching, particularly for unseen entities (Santini et al. 2026). Therefore, the proposed approach aims to supplement, rather than substitute, established hermeneutic practices by externalizing aspects of the interpretive process into a transparent, reproducible computational representation.

This goes beyond post-hoc explainability techniques in XAI, which reveal *how* a model processes input but not *which interpretive criteria* led to a specific output in a given run. By representing prompts, agents, and annotation schemas as graph nodes linked to every extracted entity, Viewsari makes the hermeneutic process, not just the result, transparent. In the field of (digital) humanities, where undocumented assertions can undermine epistemic validity (Paulheim 2016, Schimmenti et al. 2026), this approach transforms extraction pipelines into machine-actionable interpretive instruments.

Our motivation relates to a broader discourse on epistemically grounded AI. Research in epistemology has begun to investigate the conditions under which knowledge produced by generative systems can be considered valid, attributable, and contestable (Chen et al. 2026, Levin 2026). Provenance research in machine learning and data management has established formal frameworks, most notably the W3C PROV standard (Lebo et al. 2013), for documenting the derivation of computational outputs. More recently, scholarship on epistemic injustice in AI (Kay et al. 2025, Bilal et al. 2025) and on the societal consequences of opaque knowledge construction (Duede 2023) has framed the absence of interpretive documentation as an ethical issue and a missing communication with the applied science domain, rather than a pure technical gap.

Viewsari tackles all three of these layers and adopts PROV-O as the baseline for formal provenance representation, treats the generative model as a documented epistemic agent rather than a black-box machine, and foregrounds the interpretive act of each productive step as a constitutive component of the KG.

Additionally, our approach aims to surface hidden information through interpretive modeling. Although KBs include identifiers for prominent Italian Renaissance artists such as Michelangelo or Leonardo da Vinci, uncovering hidden content could challenge established canons in art production. Incorporating long-tail and uncertain sources can provide missing context and clarify how interpretive traditions have influenced the understanding of artistic movements (Ondraszek et al. 2024a). Furthermore, addressing attribution challenges in generative extraction and comprehensive modeling of epistemic processes benefits communities beyond art history and DH. The

objective is not to supplant established methodologies, but to ensure that digital representations remain traceable to their underlying reasoning, thereby supporting both hermeneutics and source criticism. This principle also applies to other fields engaged in constructing knowledge representations, including archival documents (Vafaie 2024), scientific KGs (Dessì et al. 2021, Zloch et al. 2025), experimental processes in materials science (Norouzi et al. 2024), and social network analysis (Karjus 2025). Consequently, Viewsari demonstrates broader relevance and supports its potential for transfer to other applications.

We claim that the primary methodological contribution of this dissertation is ontological rather than computational. Viewsari establishes a representational infrastructure that enables the expression of interpretive, provenance-aware KGs. This contribution encompasses not only a novel extraction algorithm but also a reproducible methodology and paradigm for knowledge representation.

1.5 Publications

Parts of the research presented in this dissertation have previously appeared in preliminary form in peer-reviewed workshops, conference contributions, and doctoral consortium presentations. They serve as documentation of the intermediate stages of conceptual development, while this dissertation substantially extends and integrates these results into a combined methodological framework in a monograph.

Early modeling approaches underlying the ontology design and KG construction were introduced in:

- S. R. Ondraszek, G. Petri, U. Blumenthal, L. Dieckmann, E. Posthumus, and H. Sack (2024a), *eXtreme Design for Ontological Engineering in the Digital Humanities with Viewsari, a Knowledge Graph of Giorgio Vasari's* The Lives
  The requirements engineering process, user stories, CQs, and iterative design phases reported in this paper are substantially expanded and formalized in Section 7.3, and their results are documented in Chapter 8.
- S. R. Ondraszek, H. Sack, and E. Posthumus (2024b), *Viewsari: New Perspectives on Historical Network Analysis in Giorgio Vasari's* The Lives *Using Knowledge Graphs*
  The cooccurrence and social network analysis introduced in this paper is contextualized as prior work in Section 5.2.4.1 and re-examined as the corpus-level baseline for the Viewsari KG construction in Section 7.2, where the prototype network graph is shown to motivate the formal ontology design.

Reusable ontology modeling patterns and abstraction strategies were further developed in this dissertation and have been previously explored in:

- S. R. Ondraszek, B. Sartini, M. van Erp, A. Poltronieri, P. Lisena, and H. Sack (2024c), *One Pattern to Express Them All? Towards Generalized Patterns for Ontology Design in the Digital Humanities*
  The comparative pattern analysis across other similar projects introduced in this paper is updated and substantially extended in Section 10.3. The Computational Provenance (COMP-PROV) pattern is mapped against the same reference projects and identified as a novel contribution unique to Viewsari.

The core extraction pipeline is presented in:

- S. R. Ondraszek, C. Möller, J. Waitelonis, and H. Sack (2026), *ObliquER: Implicit Entity Recognition and Linking in Low-Resource Historical Texts via Large Language Models*
  Under review at the ACM International Conference on Information and Knowledge Management (CIKM 2026).
  The ObliquER pipeline for implicit and explicit entity recognition and linking, including the prompt engineering strategy, the low-resource adaptation via LLMs, and the evaluation against the ground truth annotation corpus, is described in full in Chapter 7 and evaluated in Chapter 9.

Applications of related KG methodologies in adjacent domains and infrastructures are presented in:

- S. R. Ondraszek, T. Tietz, E. Posthumus, H. Fliegl, J. Waitelonis, and H. Sack (2025a), *Indexing Historical Research Data: MemO and the NFDI4Memory Knowledge Graph*
- S. R. Ondraszek, T. Tietz, J. Waitelonis, and H. Sack (2025b), *Interlinking Research Data and Services in the Historical Sciences with MemO and the NFDI4Memory Knowledge Graph*
  Both papers report on the NFDI4Memory infrastructure work that provided the context for the Viewsari transferability study.[8]
- S. R. Ondraszek, J. Waitelonis, K. Keller, C. Niessner, A. M. Jacyszyn, and H. Sack (2025c), *Ontologies in Motion: A BFO-Based Approach to Knowledge Graph Construction for*

[8] MemO (Memory Ontology) is the formal ontology developed within the NFDI4Memory consortium, the German National Research Data Infrastructure for the historically oriented humanities (`https://4memory.de`), to represent research data, source metadata, and provenance chains for historical sources. It extends NFDIcore, the shared mid-level ontology for all NFDI consortia (Bruns et al. 2024), and provides the semantic backbone of the NFDI4Memory Knowledge Graph, a FAIR-compliant shared data space linking heterogeneous historical databases through shared identifiers and ontology-based interoperability. The MemO ontology is available at `https://nfdi.fiz-karlsruhe.de/4memory/ontology/`.

Motor Performance Research Data in Sports Science
This paper demonstrates the transferability of the ontology engineering methodology to a domain outside the DH. It illustrates the broader applicability of the methodology beyond the DH domain.

An overview of the research agenda and preliminary results was presented at the Doctoral Consortium of the International Semantic Web Conference in 2025:

- S. R. Ondraszek (2025), *Viewsari: Enabling New Perspectives on the Renaissance with a Knowledge Graph of Giorgio Vasari's* The Lives
  This doctoral consortium contribution presents an early integrated overview of the research agenda, spanning the theoretical framing (Chapter 2), the Vasari case study (Chapter 6), the ontology and pipeline methodology (Chapter 7), and the preliminary results (Chapter 8). The dissertation substantially extends all of these aspects.

1.6 Dissertation outline

The rest of the doctoral dissertation follows this road map. Chapter 2–Chapter 4 provide a theoretical framework as preliminary background; they introduce the concepts of interpretation, provenance, and epistemic knowledge, and outline the technical foundations. Subsequently, Chapter 5 introduces the state of the art and examines the limits of current KE and modeling, with a specific focus on LLMs and ontologies in the DH. Furthermore, Chapter 6 presents the applied case study and details on Giorgio Vasari's work as a data foundation. Chapter 7 is the main part of this work and describes the proposed approach and the design process iterations. Chapter 8 reports the resulting artifacts, namely the final Viewsari ontology, the ObliquER extraction and linking outputs over the Viewsari dataset and ArtPedia, the populated KG, and the web interface. Chapter 9 then evaluates these contributions against formal criteria for an ontology-engineering and methodological artifact, namely logical consistency, competency-question coverage, and epistemic adequacy, complemented by a comparative modeling study against CIDOC CRM and a baseline and ablation analysis of the ObliquER pipeline. A transfer of the approach is presented in Chapter 10, which updates the comparative ontology pattern analysis across four DH projects and demonstrates the methodology's transferability through a pilot that applies the Viewsari provenance ontology and ObliquER to nineteenth-century German-language documentary sources on the Illuminati order. Chapter 11 and Chapter 12 finalize this dissertation by discussing the findings and concluding the work.

2 A bird's perspective: Theoretical and methodological foundations

The following sections present the theoretical framework, basic terms, and concepts for understanding the methodology of this dissertation. More detailed chapters of the adjacent disciplines, such as the Semantic Web, are discussed in the corresponding chapters Chapter 3 and Chapter 4. Similarly, the state-of-the-art applications of these methods are presented in Chapter 5.

In this chapter, we outline the transformation of unstructured sources into structured, machine-actionable representations to create KGs. The theoretical considerations we present here overlap with XAI (Adadi and Berrada 2018); however, our aim is to model the epistemic generation process conceptually.

2.1 The age of generative artificial intelligence

What even is the age of GenAI? Is it a new definition of the golden age, but now, instead of peace, prosperity, and a creative peak, we have war and a rise in teenagers having chatbots as friends (Zhang et al. 2025c)?

AI has certainly shaped not only scholarly endeavors. Broadly, it denotes the design of computational systems capable of performing tasks that typically require human intelligence (Russell and Norvig 2010b). Machine learning (ML), a subfield of AI, builds such models by learning statistical regularities from data; they learn from examples rather than following hand-crafted rules (Mitchell 1997). A related concern in ML is interpretability, defined as the extent to which a model's decisions can be made transparent to human users by identifying which inputs or features influenced a given prediction (Molnar 2025).

While AI has been around since approximately the 1950s (Russell and Norvig 2010a), now, we can observe a shift in qualitative capabilities and in the deployment of systems that extends beyond mere increases in computational scale and their prevalence in everyday life, such as social media communication, commerce, and science – as we can observe, for example, in the 2025 AI Index Report (Maslej et al. 2025) and former analyses of the impact of AI in everyday life (Elliott

2019). LLMs are the most famous among GenAI developments. They are language models trained on large corpora that can generate text through parametrized probabilistic inference over language (Vaswani et al. 2017, Brown et al. 2020), thus explaining the label *GenAI* (Banh and Strobel 2023). Their capabilities result from specific architectural decisions, most notably the adoption of the transformer's self-attention mechanism (Vaswani et al. 2017). In addition, scaling regimes that systematically relate parameter count, training data volume, and computational resources have been empirically established (Kaplan et al. 2020, Hoffmann et al. 2022). These models can be fine-tuned or prompted to perform a wide range of downstream tasks with minimal task-specific supervision (Brown et al. 2020, Bommasani et al. 2022). More recent work extends them into *agentic* systems that orchestrate tool calls, retrievals, and multi-step reasoning chains rather than producing a single conditioned output (Plaat et al. 2025, Zhang et al. 2025a, Liu and Shu 2026); this reframing multiplies the layers at which interpretive decisions occur, since each tool, retrieved context, and intermediate verification step is itself a conditioned generation. However, the conditions that determine a given output, including the composition of training data, the scale of the architecture, and the formulation of prompts, are documented inconsistently across the literature (Bender et al. 2021).

This development is as heavy as the term 'large' implies. It has profound epistemological consequences for knowledge-intensive disciplines because generative systems do not merely automate existing analytical tasks; they operationalize interpretation (Floridi and Chiriatti 2020, Bender et al. 2021, Bommasani et al. 2022). When an LLM resolves an ambiguous reference, assigns a type to an entity, or infers a relation between two textual mentions, it performs an act previously performed by trained human judgment, which resulted in the so-called creation of reliable ground truths (GTs) (Pustejovsky and Stubbs 2012, Wang et al. 2023a). What distinguishes this from earlier computational approaches is not the presence of interpretation, which was always there, but its *invisibility*.

In classical supervised learning, interpretive decisions were, at least to a degree, locatable. It is possible to revisit annotation guidelines, inspect the training datasets, or ask annotators for insights into their decisions – even if those are nowhere to be found in the results. The interpretation was frozen at a documented moment in the pipeline and, in principle, traceable (Russell and Norvig 2010a, Pustejovsky and Stubbs 2012, Jurafsky and Martin 2026).

With the rise of generative models, this breaks down the bridges that academia has built, and interpretation now occurs at both inference and training time (Meng et al. 2022, Luo and Specia 2024). A prompted language model does not retrieve a preformed label; it dynamically constructs an interpretation, shaped by various parameters. Two researchers conducting the same extraction task with subtly different prompts may obtain systematically divergent results with no record that the divergence occurred at all. The widespread adoption of a small number of foundation

models across humanities projects produces a different kind of flattening. Individual runs remain stochastic; the same prompt may yield different outputs on repetition, but at the level of the field, interpretive outputs cluster around the priors, training data, and inductive biases of a handful of dominant models. The variability is local; the convergence is distributional. This is not greater accuracy but the uniform, at-scale application of a single parameterized worldview (Floridi and Chiriatti 2020, Bender et al. 2021). Canon formation, as we will discuss in Section 2.3, now presents us with the following situation: the victors do not merely write history – they fine-tune the models we use daily (Mizrahi et al. 2024).

It is worth being precise about what this does *not* claim. We do not hold that human and machine interpretation differ in *kind*, nor that they call for different documentation in principle. Both are interpretive acts, and Viewsari models both with the same machinery, where each is an activity over the same annotated text span, whether the agent is a human annotator producing GT or a prompted model producing an extraction. The additional provenance that machine extraction demands therefore closes a *visibility* gap, not an ontological one; and the same documentary discipline was always owed to human interpretation, generative extraction merely making its absence impossible to ignore.

The object of study here is *IE*, the text-to-KG step, not human annotation as such. Viewsari models the human GT process because it must be documented on the same footing, but the interpretive act under analysis is the machine's.

In this context, provenance is not limited to the origin of the training data but extends to the selection and curation of training corpora, architectural design, parameterization, the training process itself, and downstream decisions at application time, including prompt formulation, sampling strategies, and temperature settings. Each of those layers encodes specific interpretive commitments that collectively shape model outputs. However, most of these commitments remain inaccessible, whether due to proprietary restrictions, inadequate documentation for humanities research, or the absence of a shared technical vocabulary among non-experts. This is not merely a theoretical concern. Nearly half of surveyed US respondents report using AI tools for information-seeking (Shao 2025), often unable to distinguish generated output from verified facts. A key reason for this difficulty is the phenomenon of hallucination and the ability of LLMs to disguise it as fact. They generate outputs that are statistically probable yet factually incorrect, thereby appearing confident. Because these systems predict the most likely next token based on patterns in training data rather than independently verifying claims, hallucinations are not anomalies but an inherent feature of the architecture (Ji et al. 2023, Shao 2025) – but how should we decipher this without knowledge of the training data and what bias might have been introduced in the pretraining? The consequences range from the tragicomic, as seen in Figure 2.1, screenshots of search overviews turned memes, with Google's AI overview suggesting users consume rocks daily, tips drawn

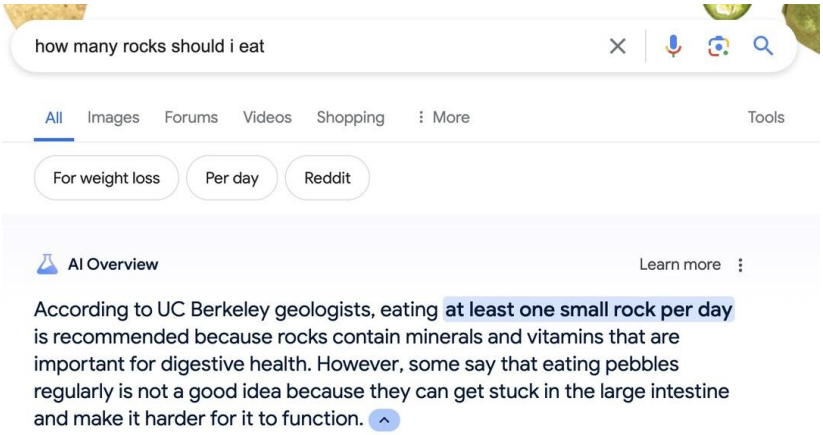


Figure 2.1: Google's AI overview responding to the query *"how many rocks should i eat"* with a confident, sourced-sounding recommendation to consume at least one small rock per day for digestive health, attributed to UC Berkeley geologists. The response draws on a satirical article from *The Onion* that the model was unable to identify as such; a widely circulated example of questionable AI output presented with the epistemic authority of a verified fact (Sarkar 2025).

from satirical sources the model could not identify as such (Williams 2024, Sarkar 2025) to the serious, including hallucinated citations in legal declarations (Zhao 2024, Shao 2025). What sounds ridiculous at first takes on an edge when a model silently assigns an interpretation to the text and no one flags it as a hallucination or falsehood.

While it is possible to identify who annotated a sentence in the Penn Treebank (Taylor et al. 2003), this is not feasible in LLMs and their generations because the interpretive act is distributed across billions of parameters influenced by training data whose composition is frequently proprietary or undocumented (Bender et al. 2021, Bommasani et al. 2022). This problem is particularly acute for historical sources, which are already interpretively dense: a sixteenth-century text such as Vasari's *The Lives*, where the canon was constructed over centuries by scholarly actors (Nova 2004, Blum 2011, Pon 2014) and carries centuries of editorial mediation and translation choices before it even reaches a language model's training corpus.

A perfect storm: the epistemological consequence is that interpretation is transformed into apparent objectivity at an unprecedented speed and scale. Generative model outputs present themselves rhetorically as retrievals, implying that an answer was found rather than constructed. However, this implication is inaccurate, and nowhere more consequentially than in the context of historical sources, where the boundary between documented fact and interpretive tradition is inherently unstable. Consequently, extracted knowledge appears as a stabilized fact rather than as a hypothesis grounded in textual evidence. GenAI risks imposing a new interpretive layer over

that canon, reconstructing the same figures while rendering their actors and choices invisible. This context motivates the theoretical framework developed in the following sections.

Prior to justifying any engineering decisions, it is essential to define the type of knowledge being modeled and the associated epistemic conditions. Section 2.2 explores the shared interpretive nature of human and machine analysis, arguing that extraction should not be understood solely as detection. Section 2.3 examines how cultural authority influences what is recognized as established historical knowledge. Section 2.4 introduces the concept of provenance as necessary for epistemic accountability. Finally, Section 2.5 considers epistemic uncertainty not as a modeling error to be minimized but as an inherent property of human knowledge that any adequate representation must preserve. Collectively, these commitments establish the theoretical foundation for the dissertation. Finally, Section 2.6 situates these commitments within the methodological trajectory of DH and articulates the cumulative gap that Viewsari addresses.

2.2 When interpretation is shared between humans and machines

Especially in historical sources, the content of interest often depends heavily on contextual understanding. To represent such knowledge, it is necessary to reconsider the nature of the statements being modeled. Humanities texts do not directly encode facts; they encode interpretations, thus epistemic findings (Gadamer 1960, McCarty 2005). This can be observed in descriptions such as "a panel for the Friars of S. Francis," which, for us, certainly signifies the mention of an artwork, but which one? The panel paintings in the church of Saint Francis in Pisa? A catalog description like this may refer to Giotto's paintings that depict Saint Francis Receiving the Stigmata.[1] However, this reference is not clear without background knowledge or supporting documentation. In real-world applications, as has been done for hundreds of years, a researcher trying to piece together which artwork might be intended there needs to collect scattered clues in a manner similar to detective work (Baxandall 1985). Identifying the delinquent – or in this case, the artwork – solely from this surface form, without fingerprints or further clues, is not possible. Recognition of a referent is an interpretation, an epistemic act, rather than one of lexical recognition. *Extraction* cannot be understood as detection alone but is grounded in textual evidence (Hovy and Lavid 2010).

[1] Wikipedia entry: `https://en.wikipedia.org/wiki/Saint_Francis_Receiving_the_Stigmata_(Giotto)`

We can state the boundary directly. A source is *interpretive* when reference resolution is not decidable from the text alone. Entities are introduced descriptively, and fixing the referent requires an inferential act. By the same criterion, a source is *documentary* when reference is canonical, and resolution is effectively a lookup, as in a born-digital catalog record, a structured database export, or an administrative file in which a name, a date, or a case number is taken *1:1* from the text. Such sources still warrant full documentation of the extraction process, since transparency is owed regardless of source type, but the act they call for is *detection*. The extractor records what the text already states. Vasari's *The Lives* sits at the opposite pole. When "a panel of Our Lady in blue and red" must be linked to Giotto's fresco "the Madonna" as a referent, recognition and identification are no longer a lookup but an interpretive act, and the separation of the textual mention from its referent becomes unavoidable. This dissertation is concerned with the latter case, where extraction is an epistemic act rather than a retrieval, though, as the administrative example shows, the documentary discipline it demands is not exclusive to that case.

In many domains of the DH, these processes are now partially automated. Large-scale digitization and the availability of computational tools have enabled analyses of corpora that could not be conducted by individual scholarly capacity alone. Automated pipelines, increasingly supported by language models, operationalize these analyses and produce statements about probabilities as extractable content, along with guidance on resolving ambiguity (McCarty 2005, Jockers 2013, Meroño-Peñuela et al. 2014, Seddone 2025, Karjus 2025). Here, this interpretive dimension becomes even more explicit, as these systems do not eliminate interpretation; they formalize it. Prompt formulations and training data distributions encode methodological choices about relevance (Zamfirescu-Pereira et al. 2023, Mizrahi et al. 2024).

In this case, the act is not a mere detection of meaning but an instantiation of a parameterized model of interpretation. This observation, however, is not exclusive to generative systems; even in non-generative supervised learning, human annotation encodes interpretive divergence in inter-annotator agreement or disagreement. Aroyo and Welty (2015) argue that this is not noise to be reconciled, but a signal of interpretation and subjective understanding that needs to be modeled as such.

2.3 Epistemology – When cultural authority constructs a canon

However, another question remains: namely, how content from historical sources is interpreted. A question of factuality: is historical knowledge in the humanities ever 'objective', mere facts?

A frequently used expression, "history is written by victors,"[2] emphasizes that most would probably reply with a 'no'. This touches on the definition of epistemology. Epistemology is the philosophical inquiry into what knowledge is, how it is formed, and when it occurs; in its traditional formulation, knowledge requires justified true belief (Saalmann 2020, Steup and Neta 2025, Mugleston et al. 2025). Reliable knowledge found in books, museums, archives, and similar sources comprises not only discovered facts but also extensive information structured through interpretive practices conducted by actors such as scholars, political regimes, and cultural institutions. Over time, these actors construct narratives that organize historical material into coherent frameworks of significance. These frameworks often manifest as a canon, a selection of works, figures, or events regarded as exemplary within a field. Consequently, canon formation involves processes of selection, interpretation, and evaluation rather than neutral description (Foucault 1980, Burke 2012). The same can be said for the age of AI (Crawford 2021, Nygren and Drimmer 2023).

In this dissertation, the chosen case study, Giorgio Vasari's *The Lives*, is an eminent example of this scenario. As Rubin Patricia Lee (1995) demonstrates, there is a method to Vasari's reproduction of stories, acting as if he were a 'neutral' historian. However, she notes that Vasari was not merely recording history and that the work, often taken as a reflection of historical facts in art history, is in fact a handpicked selection of artists whom Vasari deemed worthy of inclusion. This is also underlined by the publication of the second edition of *The Lives* in 1568, which shows revisions Vasari made between the first (1550) and second editions and documents how canonical judgments are not fixed but renegotiated over time, shaped by circumstances, new relationships, and shifting cultural values. What appears as 'factual' knowledge is, on closer examination, the outcome of repeated interpretive choices made by a situated historical actor.

In computational representations of cultural knowledge, this uncertainty underscores the importance of provenance. Documenting the source, interpretive context, and generating agent of extracted statements enables tracing the emergence and evolution of canonical claims, thereby preserving the interpretive nature of humanities knowledge.

In the age of GenAI, the problem of claims spread without connection to their original context has been extended to concern a lack of provenance information, especially in LLM-assisted research. This concerns outputs created by such models, where it is often unclear where the line is drawn between production and reproduction, especially when information is presented without a link

[2] Probably stems from the phrase "History is almost always written by the victors and conquerors and gives their view. Or, at any rate, the victors' version is given prominence and holds the field," said by Jawaharlal Nehru. The origin of this quote is not fully traceable: `https://en.wikiquote.org/wiki/Jawaharlal_Nehru#The_Discovery_of_India_.281946.29`

back to its source, since the influence is hidden in the model's parameters rather than in citable sources (Earp et al. 2025).

The main concern in the following section is therefore not the emergence of GenAI as a tool, but rather the epistemic challenge it poses by rendering the provenance of claims inescapable at scale.

2.4 Provenance and digital hermeneutics

Because interpretive statements originate from acts of reasoning, their validity depends on traceability (Mailloux 1982). Accordingly, provenance in this dissertation does not merely describe research into an object's history in the traditional art-historical or archival sense (Tompkins 2020). Rather, it concerns *the relationship between the interpretive act and the result.*

From the perspective of digital hermeneutics, algorithmic operations are themselves interpretive acts that mediate between textual evidence and structured representation (Capurro 2010).

In computer science, the term provenance often refers to the traceability of computationally executed workflows; this can include documentation of input data, results, agents, and different processing steps (Simmhan et al. 2005). With recent developments in AI, the term now also often encompasses discussions of the audit trail for LLM-generated output to better determine the sources of hallucinations (Fayyazi et al. 2025). Anchor-constrained, agentic frameworks such as the Anchor–Extraction–Verification–Supplement pipeline put this audit-trail logic into practice by tying every generated triplet element to a span the model has first identified in the source (Yang et al. 2026); the resulting record is workflow provenance in a strong sense, but the provenance it carries is the textual position of the anchor, not the interpretive convention under which the model decided that the anchor denotes what it denotes.

In humanities research, such workflow provenance needs to be extended to include not only how data were processed, but also how a particular interpretive claim was made (McCarty 2004, Gradmann and Meister 2008). When we return to the example, this means that we need to document why a system – be it human or artificial – decides that "a panel for the Friars of S. Francis," describes "St. Francis of Assisi Receiving the Stigmata."

The inaccessibility of a language model's internal representations is not a limitation to be overcome but the epistemological condition that motivates this work.

Just as a historian cannot reconstruct Vasari's cognitive process of attribution, but can document the circumstances under which he wrote – the commission, the period, the available sources – so, too, can we document the circumstances under which an LLM produces an extraction. This is what we consider the relevant provenance, not a model of interpretation itself, but a formal

account of the conditions that shaped it. The analogy to historical source criticism is deliberate, since Viewsari applies to machine-generated text the same epistemic discipline that historiography applies to human-authored text (Bernheim 1908, Droysen 1868, Capurro 2010). Source criticism as codified by Bernheim (1908) does not ask whether a source was true, but under what conditions, by whom, and to what end it was produced; the renewal of this question for digitally mediated scholarship (Fickers 2020, Schlotheuber and Bösch 2015, Weller 2013, Studies in Digital History and Hermeneutics) extends it to the database, and Viewsari carries it one step further, to the generative model whose extractions now populate that database.

This conception of provenance should, again, be distinguished from interpretable ML (Schlegel and Sattler 2025, Padovani et al. 2025). For example, asserting that Vasari describes a friendship between Andrea Verrocchio and Leonardo da Vinci involves different considerations. In a context leaning toward computational engineering, provenance may refer to identifying spans referring to Verrocchio and da Vinci as 'persons' and to identifying which model produced this output. 'Interpretation' here means an explanation of which tokens in a passage influenced the model toward the label *friendship*, a view useful for statistical explanation but irrelevant to the historical validity of the assertion. In the context of scholarly validity of the claim made above, the relevant questions are: *on what textual basis*, in *which passage*, and under *which interpretive convention* does it apply? A scholar who is unable to answer these questions has not made an interpretive claim but has produced untraceable noise. Here, provenance is not a post hoc quality check; it constitutes the condition under which an assertion qualifies as scholarship (Mailloux 1982, Ramsay 2011).

Figure 2.2 illustrates this contrast. The same source passage yields different interpretations of why the workshop entry took place, depending on which interpretation annotators apply.

This distinction becomes particularly relevant in the age of GenAI, especially when LLM-based methods are employed. Here, the extracted content is the result of parametrized interpretation shaped by prompts, model configurations, and contextual constraints. Without recording these conditions, extracted knowledge appears as a stabilized fact rather than as a hypothesis grounded in textual evidence (Paseri and Durante 2025, Mugleston et al. 2025).

This concern extends beyond inference-time configurations in GenAI approaches. A human annotator who, in constructing a training corpus, resolved the referent of "a panel for the Friars of S. Francis" as "St. Francis of Assisi Receiving the Stigmata" rather than the "Legend of St Francis" made an interpretive decision that shapes model behavior. Such annotation decisions are baked into model weights and, under current practice, are almost entirely undocumented and unrecoverable (Bommasani et al. 2022).

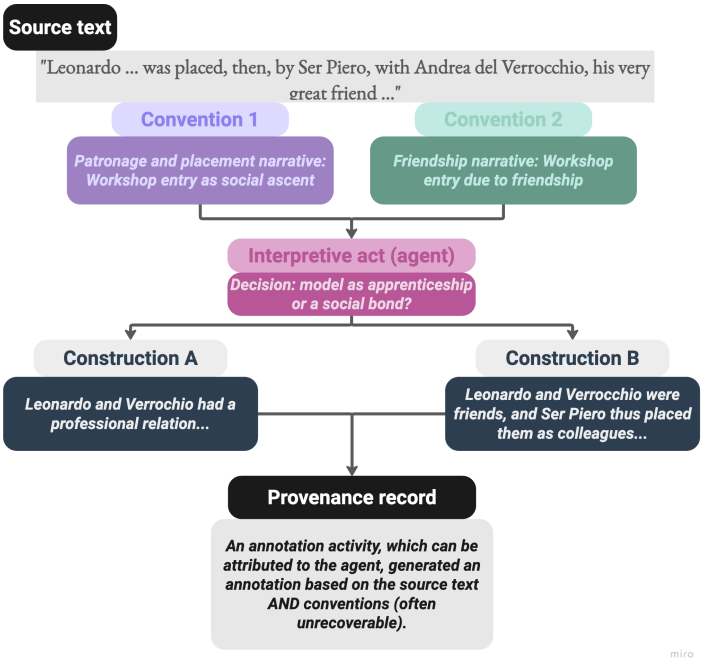


Figure 2.2: The same passage from Vasari's *Life of Leonardo* yields two distinct results depending on which interpretive convention the annotator foregrounds (Mailloux 1982); the provenance layer records the convention applied rather than resolving the ambiguity.

The subjectivity of labeling has been a recognized yet insufficiently addressed issue in NLP for over a decade. Plank (2022) and Uma et al. (2021) advocate for disagreement-preserving annotation methods that maintain the distribution of annotator judgments instead of collapsing them into a single gold label. These critiques align with epistemological concerns and make evident that the boundary between fact and interpretation must be rendered visible. Provenance tracking was not introduced as a concern by GenAI; it was merely rendered unavoidable by it (Paseri and Durante 2025, Padovani et al. 2025).

Following the specification of the provenance ontology by Lebo et al. (2013), we formally define provenance as information about the entities, activities, and agents involved in producing, delivering, or otherwise influencing a given resource, thereby enabling accountability and reproducibility. In this dissertation, provenance additionally captures *interpretive* decisions made during automated text analyses, distinguishing workflow provenance (computational steps) from hermeneutic provenance (scholarly inference).

2.5 Statistical vs. epistemic uncertainty in text analysis

Unlike, for example, observational data in experimental sciences like materials science, where instruments, experimental conditions, and measurement agents are recorded, for example, for an experiment in which iron oxidizes to Fe_2O_3 under specific temperature, oxygen partial pressure, and elapsed time (Deelman et al. 2009, Norouzi et al. 2024), statements resulting from automated text analysis in the humanities are *interpretive acts*.

The process of constructing such a statement requires the formalization of epistemic uncertainty, as illustrated in Figure 2.2. This approach is based on distinct conventions that guide interpretive acts and is fundamentally different from statistical uncertainty (Mugleston et al. 2025). Computational models represent uncertainty with probabilities or confidence scores, whereas sources in the humanities often focus on structural ambiguity in language, narrative forms, conflicting evidence, or shifting scholarly views (Baxandall 1985, Aroyo and Welty 2015, Plank 2022). In this dissertation, when discussing epistemic uncertainty, our aim is not to reduce the complexity of historical sources. Instead, we seek to make this complexity a key part of the representation of historical knowledge. Historical scholarship has long acknowledged uncertainty in the form of missing, incomplete, unclear, conflicting, or ambiguous information. In the context of DH, however, this uncertainty must be made explicit and formally represented in order to support computational processing and reasoning (Piotrowski 2019, 2026).

The growing emphasis on transparency, provenance, and reproducibility in contemporary research infrastructures underscores the need to document the interpretive origins of scholarly knowledge. Platforms such as ResearchEquals exemplify this trend by enabling the publication of not only final results but also intermediate stages of the research process, including datasets, code, protocols, and other materials.[3]

2.6 Situating the work: digital humanities and the methodological gap

The epistemological commitments developed in Section 2.1 through Section 2.5 do not arise alone. This section situates the present work within the methodological evolution of the DH and

[3] ResearchEquals web presence: `https://researchequals.com/`

articulates the gap that the Viewsari approach addresses. A detailed survey of the current state of the art follows in Chapter 5.

2.6.1 A brief overview of the digital humanities

DH is an interdisciplinary field encompassing literary studies, linguistics, art history, and the performing arts, sharing the aim of advancing humanities research and engaging with CH through computational methods (Schreibman 2004, Burdick et al. 2012). Its origins are commonly traced to the Index Thomisticus of Jesuit scholar Roberto Busa in the 1940s, an early demonstration that large textual collections become tractable to systematic analysis once organized and indexed computationally (Busa 1980). Corpus linguistics extended this paradigm by establishing quantitative and statistical analysis of large text collections as a methodological standard, shifting attention from individual works to distributional patterns across corpora (Desagulier 2017, Piotrowski 2012). The Text Encoding Initiative (TEI), formalized in the late 1980s, provided one of the first standardized frameworks for representing textual structure and meaning in machine-processable form (Sperberg-McQueen and Burnard 1994, Burnard 2014). This editorial dimension is itself the concern of textual scholarship and scholarly editing (Sahle 2016, Pierazzo 2015). *The Lives* exists in the 1550 and the revised, expanded 1568 *Giuntina* editions, in modern critical editions (Nova 2004, Vasari 2009), and in translations such as Gaston du C. de Vere's of 1912 that make their own lexical and referential choices.

Subsequent decades brought large-scale digitization and the construction of linked-data infrastructures. Initiatives such as the Women Writers Project (Women Writers Project 2023), the British National Corpus (BNC Consortium 2007), and Gallica (Bibliothèque nationale de France 1997) have made cultural materials accessible at an unprecedented scale, and projects like MiMoText (Röttgermann 2024) now build on these resources to construct domain-specific KBs. In the CH domain, the ArCo KG integrates detailed object, provenance, and bibliographic information on Italian cultural artifacts across institutional boundaries (Carriero et al. 2019). Digital art history has developed alongside these broader efforts, enabling computational access to art historical knowledge while remaining attentive to its interpretive depth (Oldman et al. 2016, Pugh 2019, Albers et al. 2020, Brey 2021, Ahola et al. 2025). The contemporary direction points toward integrating GenAI into these pipelines (Ahola et al. 2024, Cigliano et al. 2024, Leda Cempellin 2026), a development that motivates the epistemological framing of this dissertation.

2.6.2 Mind the methodological gap

Each phase of this development resolved a specific limitation of its predecessor while exposing a new one. Frequency analysis and distributional methods successfully identified regularities across large corpora but reduced relational meaning to statistical forms. Surfacing that textual mentions of *Michelangelo* and *Raphael* appear together across many paragraphs of Vasari's *The Lives* reveals a signal of connection, but discloses nothing about whether that connection is one of patronage, rivalry, emulation, or shared geography. NER and relation extraction addressed this deficit by recovering not only statistical patterns but typed entities and relations, yet they did so under the assumption that textual mentions indicate concrete referents that are explicit and surface-recoverable, an assumption that fails for the implicit, descriptive, and coreferential mentions characteristic of historical narratives (Hosseini 2019, Schimmenti et al. 2025). KG construction extended the representational scope further, mapping typed relations into ontologically grounded structures, yet the resulting triples flatten epistemic frames, so that categorical distinction between a documented assertion and an interpretive hypothesis is collapsed into the same syntactic form (Hogan et al. 2021, Schimmenti et al. 2026). We refer to this assertion-status as the PIEA, as introduced in the introduction and further contextualized in the following subsections.

Despite its might, GenAI has not resolved this gap. As argued in Section 2.1, the conditions under which a generative system produces an extracted statement, namely training data composition, prompt formulation, and sampling configuration, are themselves interpretive commitments, and they remain inaccessible under current practice (Bender et al. 2021, Bommasani et al. 2022). Each successive computational layer has expanded what can be represented while leaving the conditions of its production progressively less traceable.

2.7 From unstructured to structured textual (historical) sources

As can be taken from the theoretical foregrounding of Section 2.1 to Section 2.5, the focus in this dissertation is on textual historical sources. These textual sources are fundamental to research across diverse domains of the humanities, including archival records, letters, biographical writings, and literary texts (Flanders and Jannidis 2019).

It is important to distinguish among different types of sources, particularly because they vary significantly in the degree to which their content is explicitly structured. This subsection introduces this distinction at the outset, as we focus on the process of transforming unstructured

textual (historical) sources into structured forms. What, then, defines a source as structured or unstructured?

Essentially, spreadsheets and database records are *structured*. Fields are predefined, values typed, and meaning largely self-contained within the schema. There are also semi-structured sources, such as XML documents or metadata records, which carry their own organizational markup but still require interpretation to be useful (Flanders and Jannidis 2019).

Unstructured sources include natural language texts, which are the primary focus of this dissertation. In these cases, no schema imposes meaning on the content (Jurafsky and Martin 2026). Instead, meaning is embedded in syntax, rhetoric, and cultural or discursive context, which are not inherently captured by any predefined data model. And, in addition, historical texts constitute a particularly complex subset of unstructured sources. They not only lack explicit structure but also carry cultural specificity, having been written in particular historical moments with conventions of naming, allusion, and attribution that may no longer be transparent (Baxandall 1985, McCarty 2005). The entities referenced are often unnamed, the relationships asserted are frequently implicit, and the claims made are sometimes hedged, ironic, or contested.

So, the culturally valuable content is often hidden, lost (until discovered), in context-dependent structures. Accordingly, the digitization of documents alone does not make the content of such sources machine-readable or processable.[4] Digitalization, defined as the broader restructuring of social and epistemic practices around digital infrastructures (Brennen and Kreiss 2016), imposes similar demands on the humanities. Knowledge acquisition processes that previously relied on analog scholarly conventions must now be adapted to digital environments. However, this transition is complex because the knowledge in historical texts often resists the structured representations required by computational systems.

To illustrate this more concretely, consider a short passage from an art historical text, printed in a book from 1568:

> "the minister of culture to the Medici court in Florence blessed the Pope with a gold-crested angel in the chapel ceiling."

This sentence lacks explicit identifiers or names. The phrase "Florentine master" does not specify an individual; the "Pope" is not identified by era or name; the "chapel" remains unlocated; and it is unclear whether "a gold-crested angel in the chapel ceiling" refers to a painting, a fresco, or a sculpture, and whether it depicts an angel or merely features one as a motif is left open.

[4] Since we are working with a text that is already available in digital form, we will omit a full definition of digitization. More information on converting analog sources to digital form can be found in guidelines on automatic text recognition, such as (Rehbein 2017).

Extracting structured knowledge from this sentence requires interpretive decisions and contextual understanding that go far beyond basic syntactic parsing or digital encoding.

In an ideal world setting, someone has done the job and resolved ambiguities across this sentence; then, a relational database representation might record the result of those decisions as follows, making the knowledge from the data machine-actionable :

Table 2.1: Example of a structured relational representation of artworks.

Artwork_ID	**Title**	**Creator**	**Location**
A001	Gold-Crested Angel	Giorgio Vasari	Chapel Ceiling
A002	Scenes from St Francis	Giotto	Church of Saint Francis

But – where did the interpretive decisions go, and how do we know, for certain, that the descriptive phrase "the minister of culture to the Medici court in Florence" refers to Giorgio Vasari? Linking this to a specific individual requires contextual inference rather than lexical recognition of a surface form. This interpretive character cannot be captured by a straightforward transformation into structured data; in humanities texts, such a transformation is rarely detected in isolation but rather requires scholarly inference grounded in contextual reasoning. This epistemic challenge constitutes the central problem addressed in this dissertation. How can interpretive textual knowledge be represented computationally without diminishing its inherent ambiguity and contextual richness, flattening all steps of the in-between?

The transformation from raw text to structured knowledge is commonly framed as a progression from data to information to knowledge to wisdom; in humanities contexts, this progression is non-linear, since converting text into analyzable data is itself an interpretive act (Floridi 2013, Zeng 2017, Picca 2020, Hyvönen 2026). Language models, in contrast to rule-based methods, manage the variability and implicitness of unstructured text (Jurafsky and Martin 2026), but introduce epistemic opacity at each step.

Motivational aspects for the transformation of unstructured (historical) sources. The motivation behind this transformational process from plain text to knowledge or wisdom is manifold, including enabling large-scale analysis, supporting information retrieval, facilitating cross-source comparison, and enabling computational reasoning. See, for example, Figure 2.3: If someone is looking for works on the Renaissance painter Michelangelo and his workshop in the Renaissance, a system must be able to retrieve sources containing material on that. Searching manually for patterns of "with whom did Michelangelo collaborate and when?" across thousands of documents is infeasible without machine-actionable representations, especially when only raw

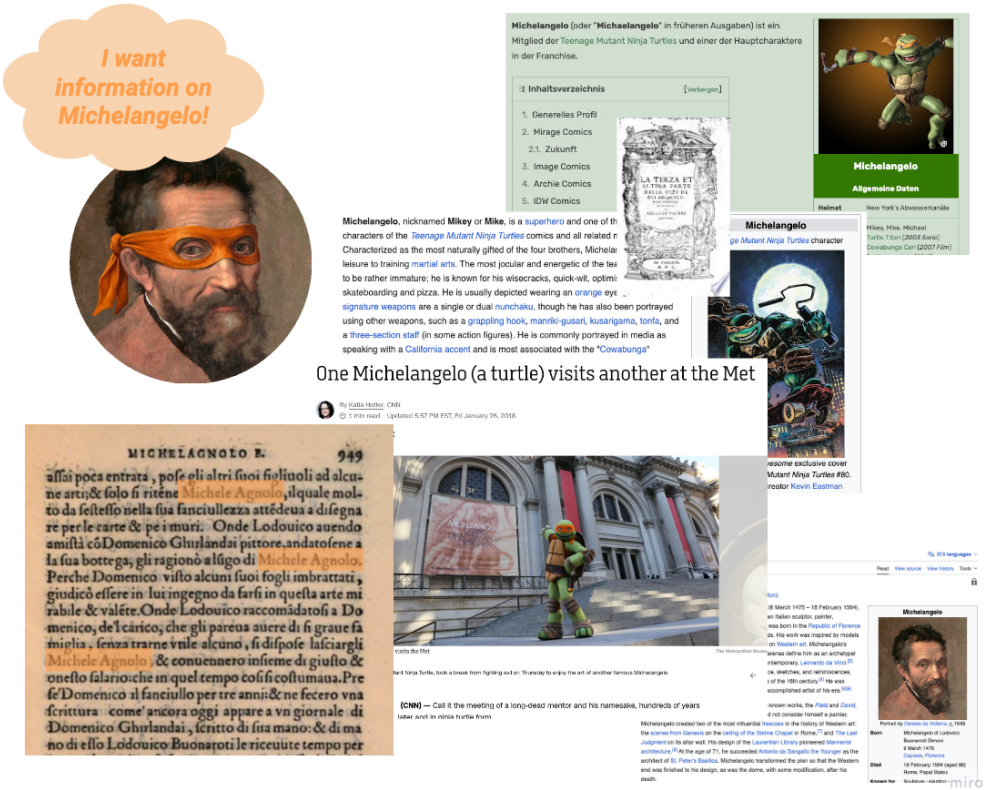


Figure 2.3: The ambiguity problem illustrated: a search for "Michelangelo" may retrieve results relating to the Teenage Mutant Ninja Turtle of that name rather than Michelangelo Buonarroti, the Renaissance sculptor and painter, and references to him in Giorgio Vasari's *The Lives*. Resolving such ambiguities requires EL grounded in contextual knowledge – which is only possible when textual sources are transformed into structured representations and the entities are modeled accordingly (see Chapter 7).

text is given (Moretti 2013, Jockers 2013, Jiang et al. 2025), so, ultimately, we do not want to know too much about the Teenage Mutant Ninja Turtles (though that is a matter for the EL part in Chapter 7), but about Michelangelo Buonarroti, the Renaissance artist. So, this means it is only possible if we provide an adequate conceptual representation of Michelangelo Buonarroti, as well as textual representations referring to him. Furthermore, interoperability across data collections depends on shared vocabularies and linkable identifiers. CH materials are often dispersed across repositories, even though their content might be related or even dependent on one another; structured representations enable cross-connections. A structured representation also provides a baseline for integrating heterogeneous sources and ensuring comparability within a single corpus (Flanders and Jannidis 2019). Without formal representation, digital artifacts

remain searchable collections of text rather than interconnected knowledge systems (Hitzler et al. 2010, Meroño-Peñuela et al. 2014, Wilkinson et al. 2016).

2.8 Moving towards structured representations of historical sources

According to Feigenbaum (1977), knowledge engineering is the process of acquiring, representing, and formalizing domain knowledge for computational systems. This includes methods for building knowledge-based systems (Studer et al. 1998). A key part of knowledge engineering is IE, which uses computational methods to convert unstructured or semi-structured natural-language text into structured data for KBs (Grishman 1997, Sarawagi 2008, Xu et al. 2024). The main goal is to break down text into entities and relations, which are the basic building blocks of structured knowledge (Zhao et al. 2024, Jiang et al. 2025).

A detailed introduction to the foundations of entities in the context of semantic technologies follows in Chapter 3

2.8.1 Information extraction for creating knowledge representation

IE turns raw text into structured knowledge through a sequence of steps – preprocessing, NER, coreference resolution, EL, and relation extraction – that produce the triples from which a KG is built (Nadeau and Sekine 2007, Josifoski et al. 2022, Liu et al. 2024, Zheng et al. 2024). Each step is defined in detail in Chapter 4.

KE is another approach that creates knowledge from both structured and unstructured sources. The results need to be machine-readable, machine-interpretable, and able to support inference. While KE uses some of the same methods as IE, it differs in that it not only creates structured information but also reuses existing formal knowledge, such as ontologies (Pan et al. 2024, Dessì et al. 2021).

2.8.2 Knowledge graphs and ontologies in the context of unstructured sources

This transformation is fundamental to AI tasks, which depend on actionable knowledge representations. Specifically, it pertains to populating KGs or bases with structured knowledge derived from extracting $(subject, relation, object)$ triplets from natural language (Josifoski et al. 2022). These triples lack inherent interpretability. To enable inference and interoperability, both the schema governing the domain and the assertions populating it must be formally defined. A KB, as a structured repository of formally represented facts, comprises two components: the terminological box (TBox), which defines the schema or ontology of classes, properties, and axioms governing the domain, and the assertional box (ABox), which contains statements about specific instances and their relations that instantiate the schema (Guarino and Giaretta 1995, Baader et al. 2003, Hitzler et al. 2010).

These structured triples presuppose a formal schema. Following Gruber (1993) and Studer et al. (1998), an ontology is a formal, explicit specification of a shared conceptualization.

Correspondingly, a KG is a KB in which the ABox is represented as a directed, labeled graph, with mandatory ontological grounding for every node and edge (Hitzler et al. 2010, Hogan et al. 2021).

The progression of these methods is epistemological, not merely technical, as each step adds abstraction and depends on interpretive commitment, especially when ontologies are introduced. This need for interpretation is clear in the humanities. If references are implicit or only described, extraction cannot rely solely on recognizing surface words. Instead, it has to interpret the text using a conceptual model of the subject. For example, we would recognize that Vasari's gold-crested angel is an artwork by considering the full context of the sentence. But would an NER system see it the same way? If this kind of reference comes up again, could the system recognize the connection?

2.9 The forms of entities in textual sources

If there were no intermediate representation layer, tasks such as reasoning, querying, or semantic search would have to work directly with raw text. IE creates a structure that enables the construction and maintenance of KGs. In this section, we highlight implicitness, namely implicit entities, as a key theme in this dissertation.

The central currency of IE is the entity, defined as a distinct unit of reference in a domain (Auer et al. 2007, Hogan et al. 2021). Standard pipelines assume that such units appear as surface strings recoverable through NER and linkable to a KB through EL, the technical detail of which we defer to Chapter 4. Narrative sources like *The Lives*, however, often leave entities unnamed. These are called implicit entities. Recognizing and resolving them requires context, domain knowledge, and an understanding of discourse structure, since they do not appear as standard identifiers (Hosseini 2019, Stramiglio et al. 2025).

For example, consider the sentence we discussed earlier. It includes four possible entities: a person (the minister of culture to the Medici court in Florence), another person (the Pope), an artwork (a gold-crested angel in the chapel ceiling), and an architectural feature (the chapel ceiling).[5] None of these is named explicitly. An equivalent sentence with explicit entities might read: "Giorgio Vasari presented Pope Julius III with *The Gold-Crested Angel* for the Palazzo della Cancelleria." Explicit entities can be identified by their names in the text, given that they are textually manifested. In contrast, implicit references lack clear names, often require background knowledge to understand, and may refer to entities not present in external databases – they are contextually recoverable only through domain knowledge or interpretive inference applied beyond the boundaries of the text itself. This makes them harder to process, as they blur the line between finding and interpreting entities and often need information from several sentences. Recent studies on implicit communication in LLMs further demonstrate that meaning is frequently conveyed without explicit lexicalization, requiring contextual inference and pragmatic reasoning (Tint et al. 2026). These findings reinforce the claim that computational systems must move beyond literal extraction toward modeling the interpretive reconstruction of meaning.

Implicit references can resemble coreferences. These refer to the opposite problem: when a referring expression is present in the text but does not introduce a new entity, only referring back (or forward) to an entity already introduced elsewhere. These entities' coreferents can, of course, also be explicit or implicit. Most coreferents are pronouns, epithets, shortened names, and definite descriptions ("the painter," "he," "the minister of culture to the Medici court in Florence"); all co-referential when they resolve to a previously named individual. An implicit entity may lack a specific surface form but has a real-world referent that must be inferred. In contrast, a co-referential reference includes a referring expression but no new referent and therefore must be resolved to an existing entity (Perera 2016). Another resemblance is given in the form of generic

[5] What allows us to call "a gold-crested angel in the chapel ceiling" an artwork at all is itself a classificatory commitment. The Getty's *Categories for the Description of Works of Art encodes this commitment in its Object/Work-Type category, which establishes the typology under which an item is recognized as a work of art (e.g., painting, fresco, sculpture, architectural element) before any further descriptive attribute is recorded (Harpring 2010). See* https://www.getty.edu/publications/categories-description-works-art/categories/object-architecture-group/2/.

mentions – surface forms which look like an implicit entity, like "these artworks" or "his many drawing," but refer to artworks or terms of art history more globally.

To address implicit entities, we need more than just a sequence labeler that works on surface forms. We need a system that uses a formal domain model to define valid entities and can show the reasoning steps used to resolve implicit references; we elaborate on this in the context of text-to-KG pipelines, which offer exactly this approach.

Definition: The pre-interpretive entity assumption (PIEA). The PIEA is the modeling commitment that the resolution of a textual mention to its real-world referent can be treated as completed prior to representation in a KG – that is, that an extraction system's task is to recover a referent already determined, rather than to perform the determination. Under PIEA, the categorical distinction between mention (a textual occurrence of a referring expression) and referent (the entity, if any, denoted by that mention) is collapsed in the schema: the extracted node is the referent, and the mention is metadata at most.

PIEA is the unstated default in standard knowledge representation textbooks (Hitzler et al. 2010, Hogan et al. 2021), in schema-driven extraction benchmarks (Josifoski et al. 2022, Mihindukulasooriya et al. 2023), and in the de facto standard CH vocabulary (Doerr 2003). The assumption is coherent for sources in which the referent is explicit, canonical naming is the norm, and the KB has comprehensive coverage of the relevant domain. For the interpretive sources central to this dissertation, none of these conditions holds: a substantial proportion of references are implicit and descriptive, and mention-to-referent resolution is the representational work of the extraction itself rather than its precondition. We name PIEA explicitly because the assumption is built implicitly into evaluation regimes and modeling conventions in practice but is rarely stated as such. For instance, it is observable in deployed humanities KGs such as the Odeuropa project, which collapses the textual mention and the olfactory phenomenon it reports into a single node (Figure 10.3), materializing the referent directly from the mention in exactly the manner PIEA describes. As shown in prior work, the Vasari corpus by Santini et al. (2022) resolves referents with 0% extraction provenance. On the referent axis, widely used in-KB metrics such as the GERBIL setting (Usbeck et al. 2015) evaluate only mentions whose gold entity is present in the target KB, thereby excluding OOKB mentions. This approach presupposes resolvability, which interpretive sources often lack. On the mention axis, this setting provides the system with gold mention spans, reducing the task to referent assignment. NIL-aware entity linking, which clusters unlinkable OOKB mentions into distinct entities rather than discarding them, is the deliberate exception (Blissett and Ji 2019); that this remains a comparatively neglected subtask within entity discovery and linking confirms that PIEA is the unmarked default enforced by different

representations. Our suggested three-layer architecture, introduced in Chapter 7, serves as the representational response.

2.10 Bringing the worlds together: text-to-knowledge-graph pipelines

The progression from raw text to relationally structured assertions is commonly described as a text-to-KG pipeline. These approaches integrate extraction results into an ontologically grounded representation, in which extracted information is constrained by a domain model. Knowledge engineering thus precedes and conditions the construction of such pipelines (Kertkeidkachorn and Ichise 2017, Hogan et al. 2021, Pan et al. 2024, Melnyk et al. 2022).

In other words, extraction produces structured statements; ontology-guided graph construction determines their meaning within a semantic framework. In the humanities, though, this process is more than just a technical task (Schimmenti et al. 2026, Gangemi et al. 2025).

Unlike fields with clear terminology and well-defined entities, texts in the humanities often convey meaning through narrative, rhetorical choices, and interpretation. Schimmenti et al. (2026) shows that CH texts about disputed artifacts require representations that track provenance and accommodate multiple perspectives. For example, Gangemi et al. (2024) uses a text-to-KG pipeline on a multilingual collection of musical heritage texts, showing exactly this interpretive process; every modeling choice, from selecting text spans to defining entities and designing the ontology, involves interpretation.

LLMs have become central to every part of this pipeline. These are now considered the most advanced methods, which we will discuss further in Section 5.2.4.2.

3 The Semantic Web: historical development and foundations

Knowledge representation formalizes content, while extraction technologies transform textual material into structured assertions. The following sections introduce the basic terminology and concepts necessary to understand how knowledge representations and semantic technologies work together, particularly regarding ontologies and KGs. A deeper examination of state-of-the-art methodologies follows in Chapter 5.

The Semantic Web, introduced in the early 2000s, aims to transform the web from a collection of human-readable documents into a network of machine-interpretable data (Berners-Lee et al. 2001). The Semantic Web stack, depicted in Figure 3.1, is a layered architecture of W3C-standardized technologies that convert the Web from human-readable documents into logically structured, machine-interpretable data (Berners-Lee et al. 2001, Hitzler et al. 2010). Each layer builds upon the preceding one, increasing expressivity and reasoning capabilities, and is organized into identification, data, schema, query, and reasoning layers. We use this visualization to explain the different aspects of semantic technologies, starting at the bottom and working up to the top, where logic is applied.

3.1 The World Wide Web

The Semantic Web is built on a few key web technologies. Uniform Resource Identifiers (URIs) provide global identification, the Hypertext Transfer Protocol (HTTP) handles communication, and Unicode standardizes character encoding. Together, these make sure resources are easy to find and access across different systems.

Building on this web foundation, the W3C introduced the Resource Description Framework (RDF) in 1999 (Klyne and Carroll 2004). RDF uses a graph-based data model in which information is represented as triples consisting of a subject, a predicate, and an object. Each triple makes a statement about the world:

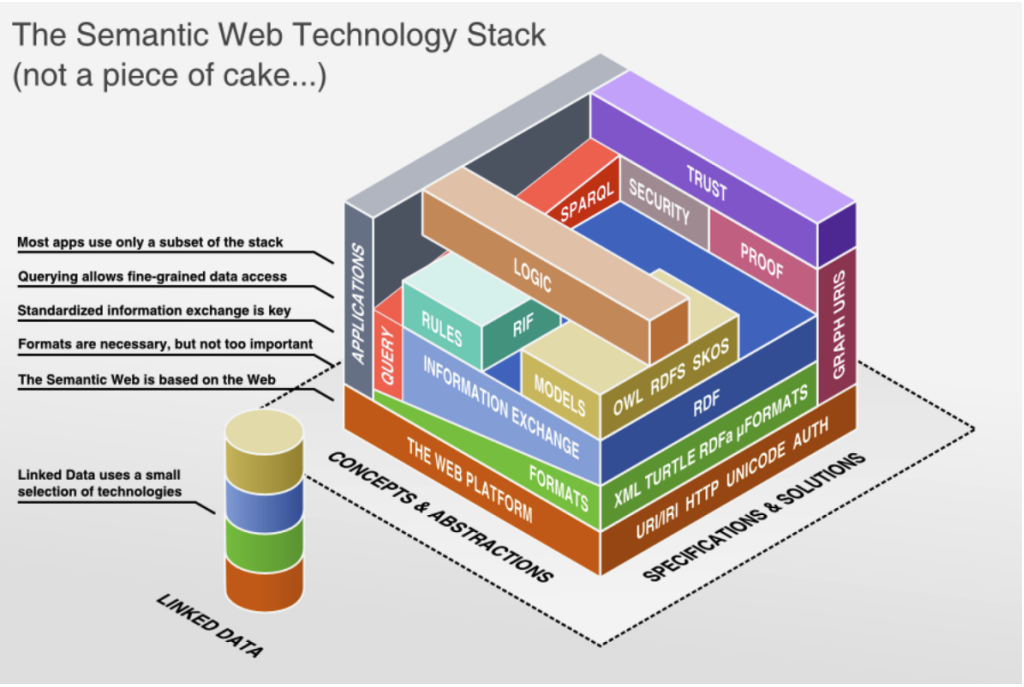


Figure 3.1: The Semantic Web Technology Stack (after Nowack 2009), illustrating the layered architecture of W3C-standardized technologies that underpin the Semantic Web. From bottom to top: identification (URI/IRI), data exchange (XML, RDF), schema and ontology (RDFS, OWL, SKOS), rules (RIF), querying (SPARQL), and higher-level layers for logic, proof, trust, and security.

Giorgio Vasari – created – Gold-crested Angel.

A finite set of RDF triples constitutes an RDF graph. These triples form labeled graphs, enabling the representation of interconnected knowledge structures rather than isolated data entries.

3.2 Ontologies and ontology languages in theory

RDF defines the structural model, but it does not provide the specific vocabulary used within that structure. This is where data models like RDF Schema (RDFS), which extends RDF to define classes and properties (Brickley and Guha 2004), or ontologies written in the Web Ontology Language (OWL), a set of knowledge representation languages standardized in 2004 (McGuinness and van Harmelen 2004), come in. These languages introduce classes, properties, hierarchies, and logical rules that describe the kinds of entities that exist in a domain and how they relate to one another. Ontologies give systems formal semantics. For example, they help systems

distinguish between a *Person* and an *Artwork*, allowing specific relationships such as *created* or *participated in*. OWL, furthermore, supports logical inference; in that case, the system can draw new conclusions from information already stated via the ontology.

OWL is not a monolithic language but a family of sub-languages with formally specified expressivity in description logics (DL), which are formal knowledge representation languages (Baader et al. 2003). Among the normative OWL 2 species relevant to applied knowledge engineering, we focus on OWL 2 DL. It supports nominals, inverse properties, role chains, and qualified cardinality restrictions while guaranteeing decidability of reasoning (Motik et al. 2012). In the Viewsari ontology (Chapter 7), we rely in particular on inverse property axioms and qualified cardinality restrictions.

It is important to distinguish between `rdfs:Class` / `rdfs:subClassOf` and `owl:Class` / `owl:equivalentClass`. RDFS defines schema-level semantics for subclass hierarchies and domain or range declarations, but it does not use closed-world or disjointness reasoning. OWL builds on this by using the open-world assumption, which means that any fact not explicitly stated or derived is simply unknown rather than true or false. This is especially relevant for humanities KGs, where incomplete knowledge is a normal part of the structure (Allemang and Hendler 2011). Using `owl:Class` instead of `rdfs:Class`, an OWL reasoner can classify, check consistency, and infer properties in the graph.

Ontologies are structured into a hierarchy of abstraction levels according to their scope and degree of domain specificity (Guarino 1997, d'Aquin and Noy 2012, von Rosing et al. 2015). Upper or top-level ontologies occupy the highest tier, as they are not associated with any particular discipline. Their categories, including classes and properties, are intended to be applicable across all domains of knowledge. Notable examples of top-level ontologies are the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (Gangemi et al. 2003) and the Basic Formal Ontology (BFO) (Otte et al. 2022).

Mid-level ontologies refine abstract concepts to enhance their applicability across multiple disciplines, without concentrating on a single domain. The CIDOC CRM (Doerr 2003) exemplifies this category. This model provides a framework for documenting CH by linking top-level concepts to specific collection data. It is utilized in disciplines such as archaeology, museology, and art history. Domain ontologies address the concepts, properties, and relationships that are specific to a particular field, such as biomedicine, legal studies, or art history. These ontologies frequently incorporate or align with upper- and mid-level ontologies to facilitate interoperability among diverse systems. Application ontologies represent the most detailed level. They extend domain ontologies by incorporating specific rules, examples, and vocabulary required for particular systems or tasks (Guarino 1997).

We see the Viewsari ontology as sitting between the domain and application levels. It represents the structure of art-historical knowledge and also includes the extraction and provenance requirements of the ObliquER pipeline.

Statements about statements – reification in ontologies Adding metadata to a triple needs extra steps, such as RDF reification (a standard since RDF 1.0). This process breaks a triple $\langle s, p, o \rangle$ into four triples using `rdf:Statement`, `rdf:subject`, `rdf:predicate`, and `rdf:object`. This method is wordy, does not formally link back to the original triple, and most triple stores do not index it by default (Hayes and Patel-Schneider 2014). To address this, named graphs were created. They link a group of triples to a graph identifier g, making a quad $\langle s, p, o, g \rangle$. Named graphs work well in SPARQL 1.1 and triple stores and are often used for dataset-level provenance. Still, they cannot add metadata to a single triple without creating a new graph for each triple. RDF-star is the latest addition and will become part of the RDF framework in RDF 1.2. It lets a triple act as the subject or object of another triple, shown as $\langle\langle s, p, o \rangle\rangle$ for an embedded triple (Hartig 2017, Arndt et al. 2022).

The choice to use reification with the `viewsari:Cooccurrence` class in the Viewsari ontology is explained and reviewed in Section 7.3.

3.3 Basics of ontology engineering

Ontology engineering methods help structure collaborative modeling. They focus on refining models step by step, reusing design patterns, and ensuring the work aligns with CQs rather than sticking to a fixed schema (Blomqvist et al. 2016). A key part of this process is the idea of a CQ. A CQ is a question in everyday language that an ontology should be able to answer based on the facts it contains (Grüninger and Fox 1995). In ontology engineering, CQs serve as requirements to guide design and, in subsequent steps, as evaluation criteria to verify that the ontology covers the relevant knowledge. They count as formally satisfied when a SPARQL query against a KG based on the ontology returns a non-empty, correctly typed set of answers.

CQs are elicited from domain experts and operationalize the scope. Instead of staying abstract, this makes it possible to model formalizations based on real-world needs from experts through questions such as *"Which artists collaborated with Vasari?"*, *"Which artworks are mentioned in more than one biography?"*, *"What provenance information is available for a given extracted entity?"*. With these questions comes a set of classes, properties, and related axioms for the ontology to define so that the corresponding SPARQL query is answerable. This method of designing ontological formalizations grounds the decisions in use cases rather than in abstract,

often merely theoretical completeness. It ensures that the resulting model is actually used and can be evaluated against concrete requirements and data.

Correspondingly, in the XD methodology (Presutti et al. 2009), CQs serve as the primary input to each development iteration, during which modular components of the ontology are released only after all associated CQs are answerable on representative test data. The CQs developed for the Viewsari ontology are presented in full in Chapter 7. The Semantic Web stack further envisions rule languages and logic layers that enable reasoning and inference. Reasoning in our work refers to the process performed by a reasoning engine (reasoner), which infers logical consequences from a set of asserted facts or axioms (Motik et al. 2012, Eiter et al. 2006, Glimm et al. 2014). Inference is the result; the derived new knowledge and relations from said process (Ulicny et al. 2014).

3.4 Basic criteria for ontology evaluation

Ontology evaluation primarily assesses conceptual adequacy, logical soundness, and fitness for purpose, all of which are qualitative metrics. Nevertheless, quantitative methods also exist to assess ontology quality, which we will explain in the following section. As we mentioned before, one central criterion of this evaluation is the use of CQs to assess fitness for purpose, but they can also be used for quantitative evaluation.

Among the qualitative desiderata for ontology evaluation, Gruber (1993) have proposed a focus on clarity of definitions, coherence of axioms, logical consistency, and extendibility, so that new specializations can be added without changing existing definitions. They also recommend minimizing encoding bias, meaning the concepts should not rely on a specific symbol-level encoding, and keeping ontological commitment minimal, so the ontology only makes necessary claims to support its intended use. These criteria are used in the design review of Viewsari in Chapter 9. Additionally, as we have introduced before, logical consistency can be verified, e.g., using OWL reasoners such as HermiT (Glimm et al. 2014).

Frameworks have been introduced to harmonize different evaluation steps, including OntoClean (Guarino and Welty 2002). These tools can be used to evaluate according to criteria such as rigidity (whether a property is essential to an instance across all possible worlds), identity (whether a property provides criteria for distinguishing instances), and dependence (whether instances require the existence of other entities). Although OntoClean is primarily suited for foundational and upper-level ontologies, it can also be used to evaluate domain and application ontologies with respect to class identity and rigidity. Commonly used is also the OOPS! (OntOlogy Pitfall Scanner) (Poveda-Villalón et al. 2014) for detecting common ontology modeling

pitfalls, i.e., properties without inverse axioms, missing domain and range declarations, or modeling relationships as attributes instead of object properties. Task-based evaluation can enhance these metrics by assessing whether an ontology supports the queries and inferences required by its specified use cases.

3.5 Core principles of querying

SPARQL (SPARQL Protocol and RDF Query Language) is the standardized query language for RDF graphs (Beckett et al. 2014, Harris and Seaborne 2013). It enables executing queries on graph structures via pattern matching, aggregation, and federated queries across endpoints.

The primary query form used in this dissertation is SPARQL SELECT. A SELECT query is a tuple $Q = \langle W,\, DS,\, P,\, M \rangle$, where:

- W is a set of projected variables (the *result clause*),
- DS is the target dataset,
- P is a *graph pattern* – a recursive structure built from triple patterns $(s,\, p,\, o)$ in which any element may be replaced by a variable $?x$, combined via `AND`, `OPTIONAL`, `UNION`, and `FILTER` operators,
- M is a set of solution modifiers (`ORDER BY`, `LIMIT`, `OFFSET`, `DISTINCT`).

Moreover, we use said queries to evaluate CQs, following Presutti et al. (2009).

3.6 The foundations of knowledge graphs

KGs are concrete implementations of the aforementioned theoretical principles of the Semantic Web. The term, first popularized by Google in 2012 to refer to a large-scale graph of entities and their relations used to enhance search results (Singhal 2012), denotes a form of graph-based knowledge representation. Similar concepts have existed for decades under various names, including semantic networks (Quillian 1967), meaning that the Google Knowledge Graph essentially consolidated the concept and sparked significant academic and industrial interest. Following the growing interest in the technologies behind the Google Knowledge Graph, various initiatives have led to the development of large, open KGs such as Wikidata (Vrandečić and Krötzsch 2014) and DBpedia (Auer et al. 2007), alongside formal definitions and literature that continue to characterize the field of semantic technologies (Hogan et al. 2021).

There are several ways to build KGs, such as manual curation, automated extraction, and schema alignment. Usually, construction pipelines use a text-to-KG method (Section 2.10), where entities and relationships are extracted from source documents, classified according to an ontology, and represented as RDF triples. This process includes NER, EL, and relation extraction. Each step involves choices that shape the final graph's structure. Domain-specific KGs, such as those used in CH, often need to be aligned with established reference ontologies or other data schemas. They also need to work with endpoint outputs such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Lagoze et al. 2002, Van de Sompel et al. 2004), which is a protocol for collecting metadata records from digital repositories, and MARCXML (Library of Congress 2002), an XML schema based on the MARC 21 standard for cataloging bibliographic metadata. This alignment helps ensure that collections can be connected. RDF triples are serialized in standardized formats, the most common being Turtle (Terse RDF Triple Language) (Beckett et al. 2014), RDF/XML (Gandon and Schreiber 2014), JSON-LD (Sporny et al. 2014), and N-Triples (Carothers and Seaborne 2014). Although they vary in readability and interoperability, these formats are semantically equivalent. Resulting triples can be stored in graph databases, also known as triple stores (or RDF stores) (Shimpi and Chaudhari 2012). Notable examples include Virtuoso (Erling and Mikhailov 2009) and GraphDB (Ontotext AD 2024). Unlike relational databases that organize data in fixed tabular schemas, triple stores offer schema flexibility, allowing the introduction of new entity types and properties without restructuring the underlying storage. This capability is particularly relevant to the Viewsari KG, where triples must remain traceable to their extraction process. We use GraphDB as a triple store, which we explain in more detail in Chapter 7.

KGs can be queried using SPARQL, matching patterns against stored triples to retrieve variable bindings corresponding to entities and relations that satisfy the pattern. For example, to retrieve all artworks attributed to Giorgio Vasari in the Viewsari KG, the following query is used:

```
SELECT ?artwork WHERE {
?artwork rdf:type example:Artwork .
?artwork example:createdBy wd:Q128027 .
}
```

Because they are often built on OWL ontologies, automated reasoning mechanisms can be effectively applied to KGs. For example, a reasoner such as HermiT, when applied to the TBox and ABox, can infer knowledge from the ABox based on definitions in the TBox. For instance, if the ontology specifies that `example:Painter` is a subclass of `example:Artist`, and an individual is asserted as a `example:Painter`, the reasoner will automatically classify that individual as a `example:Artist` without requiring an explicit assertion. Furthermore, reasoning supports

consistency checking. If an individual satisfies two disjoint classes, the reasoner identifies a logical contradiction, thereby facilitating ontology validation.

However, the underlying assumption is that entities within the KG are explicitly defined. As discussed in Chapter 1 and Section 3.3, sources in the humanities often violate this assumption. When entities that may exhibit various forms of implicitness or historical discontinuity and are only inferable through contextual reasoning are represented using a naïve graph construction, there is a risk of reducing interpretive complexity to decontextualized triples.

4 Text to knowledge: Foundations of information and knowledge extraction

Having established a bird's-eye perspective on the core terminologies and conceptual landscape in Chapter 2, we now descend to the technical foundations that support these concepts. This chapter articulates the baseline for information and KE. The discussion begins by tracing the development of language models, the models that form the foundations for both methods, from early statistical methods to transformer-based architectures that enable large-scale extraction. In subsequent sections, we elaborate on the concrete steps of IE, followed by a focused analysis of the challenges inherent in interpretive humanities texts when these methods are applied naively. These technical and methodological foundations inform the design decisions in the Viewsari project and position its contributions within the broader discourse on KE from historical sources.

4.1 A brief overview of the evolution of language models

Modern chatbots such as ChatGPT (OpenAI 2023) and Claude (Anthropic 2024) represent the culmination of extensive research in language modeling. These models are typically transformer-based and trained on web-scale datasets with extensive computational resources (Minaee et al. 2025), and, as we call them nowadays, *the* LLMs that function as therapists, friends, or ghost-writers (Zhang et al. 2025c). Recent surveys by Zhao et al. (2025b) and Wang et al. (2025b) categorize the development of language models into four principal stages: First, a stage centering on statistical language models, with n-gram models dominating through the use of conditional probabilities and frequency counts. Although these models performed effectively on short texts, they were limited by data sparsity and over-parameterization. Later, the field shifted toward neural architectures, which introduced distributed word representations in continuous embedding spaces. During this period, recurrent neural networks and long short-term memory models became prominent, enabling improved modeling of language context beyond the capabilities of

n-gram models. A breakthrough in contemporary technologies was the transformer architecture, which replaced recurrent computation with self-attention mechanisms. Self-attention computes contextual interactions between all tokens in parallel, enabling scalable training on modern hardware infrastructures (Minaee et al. 2025). Large-scale pretraining with architectures such as BERT (Devlin et al. 2019) and GPT (Radford et al. 2018, Brown et al. 2020) established the "pre-train and fine-tune" paradigm (Zhao et al. 2025b). Models are trained on massive unlabeled corpora (unsupervised) and later adapted to downstream tasks.

Subsequent work formalized scaling to improve performance based on parameters such as model and training data size, among other components (Zhao et al. 2025b). With an increasing number of parameters fed into these architectures, there is also an improvement in quality, especially in behavioral aspects such as in-context learning and instruction-following (Zhao et al. 2025b).

The progression from n-grams to LLMs is more than a sequence of architectural advances; it also lays the foundation for contemporary information and KE. Each stage enhances extraction quality, with transformer-based language models now enabling extraction from heterogeneous, domain-specific corpora and interpretive sources, such as historical documents like Vasari's *The Lives*.

4.2 What are the steps of information extraction?

Having established language models as the foundational infrastructure, it is necessary to conceptualize IE as a structured process: the systematic transformation of unstructured natural language into machine-readable representations. While contemporary extraction pipelines increasingly assign tasks to LLMs (Ramshaw and Marcus 1995, Jurafsky and Martin 2026), traditional steps are still often necessary for IE, especially when historical sources impose distinct requirements at each stage.

Taken together, these tasks constitute the technical backbone of text-to-KG pipelines, which we introduced abstractly in Chapter 2. However, as discussed in the previous section, applying these tasks to humanities sources introduces interpretive challenges that go beyond purely syntactic or statistical processing. It should be noted that the concepts underlying these tasks – such as the assumption of explicitly named entities – have historically been shaped by domains with standardized terminology, namely, the newswire domain (Augenstein et al. 2017). Narrative and interpretive humanities sources challenge these assumptions in fundamental ways (Ehrmann et al. 2023, Schimmenti et al. 2025, Santini et al. 2026).

4.2.1 Preprocessing

Preprocessing comprises a sequence of normalization operations that transform raw textual input into structured symbolic representations suitable for computational processing. This includes sentence segmentation (division of text into sentence units), tokenization (splitting sentences into lexical tokens), lemmatization (reducing inflected forms to canonical lemmas), and optionally part-of-speech tagging or syntactic parsing (Jurafsky and Martin 2026). Formally, given a document $\mathcal{D} = \{c_1, c_2, \ldots, c_n\}$ consisting of character sequences, preprocessing in terms of tokenization produces a sequence of tokens $T = \{t_1, t_2, \ldots, t_m\}$ with optional annotations $\alpha(t_i)$. For example, the sentence "Giorgio Vasari painted the vault of the chapel." may be represented in a token sequence: `[Giorgio] [Vasari] [painted] [the] [vault] [of] [the] [chapel]`. At this stage, the output is structured data but contains no semantic interpretation beyond surface segmentation.

4.2.2 Named entity recognition

NER is a sequence labeling task in which contiguous spans of tokens are classified into semantic categories. Formally, given a token sequence T, NER assigns a label $y_i \in \mathcal{L}$ to each token (often using predefined tagging schemes), where $\mathcal{L}$ denotes entity types such as `Person`, `Location`, `Organization`, or domain-specific classes such as `Artwork` (Manning and Schütze 1999, Lample et al. 2016, Jurafsky and Martin 2026). The BIO (Beginning–Inside–Outside) tagging scheme is one of the most popular, introduced by Ramshaw and Marcus (1995), and encodes entity spans using token-level boundary markers. Each token is assigned one of three positional prefixes: `B-` marks the beginning of an entity, `I-` marks a token inside the same entity span, and `O` denotes tokens outside any entity. For example, in the same exemplary sentence,

Giorgio Vasari painted the chapel vault,

a BIO annotation may yield:

```
Giorgio   B-PER
Vasari    I-PER
painted   O
the       O
chapel    B-LOC
vault     I-LOC
.         O
```

In this dissertation, our goal is to go beyond standard entity types to those relevant to Vasari's work; accordingly, the Viewsari ontology extends the list of types with domain-specific categories such as `implicit artwork mention` and `generic mention`. More on this is described in Chapter 7.

4.2.3 Coreference resolution

Coreference resolution identifies expressions within a document that refer to the same discourse entity. Formally, it partitions a set of mentions $\mathcal{M} = \{m_1, \ldots, m_k\}$ into equivalence classes representing shared referents. For instance, in the passage:

> "Giorgio Vasari painted the vault. He completed it in 1541."

the mentions *Giorgio Vasari* and *He* are grouped into the same coreference cluster (Soon et al. 2001, Lee et al. 2017). In Vasari's *The Lives*, coreference resolution is non-trivial, as these referents often span across very long span boundaries. Additionally, there is a manifold of referent types, including, e.g., a single artist being referred to by name, pronoun, or epithet ("the divine painter"), which requires context-sensitive resolution strategies.

4.2.4 Entity linking

EL addresses the task of associating recognized mentions with canonical identifiers in a KB. Formally, given a mention m and a knowledge base $\mathcal{KB} = \{e_1, \ldots, e_n\}$, the objective is to select the most probable entity e_j such that $m \mapsto e_j$, typically subject to contextual similarity constraints. This process operationalizes the broader challenge of grounding textual references in structured representations, thereby enabling downstream interpretive and computational tasks (Wu et al. 2020). The development of EL systems originated in research on named entity disambiguation, where the primary concern was to resolve ambiguities in surface forms by leveraging lexical similarity. As distributed representations and deep learning became central to NLP, these approaches gave way to neural architectures that model contextual information more effectively (Kolitsas et al. 2018). This shift culminated in the widespread adoption of the retrieve-and-rerank paradigm: candidate entities are initially retrieved using sparse or dense similarity measures, and subsequently disambiguated through contextual encoding. The evolution of these methods reflects a broader transition from surface-level matching to context-sensitive modeling, aligning EL with advances in representation learning (Wu et al. 2020).

Subsequent research extended this to multilingual corpora, domain-specific KGs (e.g., biomedical), and dynamically evolving KBs. Recent work also addresses the detection of OOKB or NIL

(unlinkable) entities, which traditional systems often ignore (Brown et al. 2020, Agarwal et al. 2022, Zhu et al. 2023). In historical and low-resource contexts, additional challenges arise from digitization noise and high entity sparsity (Ehrmann et al. 2023). EL in humanities sources is fundamentally interpretive. References often appear as indirect descriptions, such as "the Florentine master," or references shaped by contextual attributions and linguistic variation, e.g., "the Pope". As a result, linking decisions require contextual reasoning beyond lexical similarity. For example, associating "the Florentine master" with the Wikidata (Vrandečić and Krötzsch 2014) identifier for Michelangelo Buonarroti presupposes prior knowledge about authorship, place, and narrative context. These decisions are not merely technical but constitute acts of scholarly interpretation (Perera 2016, Ehrmann et al. 2023, Schimmenti et al. 2026). Furthermore, many historical entities are absent from external KBs. Long-tail cases like the aforementioned, unspecified Pope, forming uncertain attributions, demonstrate that not every mention can or should be normalized to a global identifier.

In Viewsari, unresolvable mentions are modeled as OOKB entities and assigned local identifiers within the KG, preserving the interpretive act rather than silently discarding it. More details on this methodology follow in Chapter 7.

4.2.5 Relation extraction

Relation extraction addresses the task of identifying semantic relations among entities recognized within a textual source. This process constitutes a critical step in transforming unstructured text into structured, machine-interpretable knowledge. Formally, given entities e_i and e_j occurring in a textual context c, the task predicts a relation $r \in \mathcal{R}$ such that (e_i, r, e_j) forms a structured assertion. From the example sentence, a system might extract:

$$\textit{Giorgio Vasari} \xrightarrow{\texttt{creatorOf}} \textit{Vault of the Chapel}$$

By extending beyond entity identification, relation extraction enables the structuring of relational assertions. This yields candidate triples that serve as foundational elements for subsequent KG construction (Pawar et al. 2017).

4.3 Open and closed information extraction

Closed IE is the process of extracting structured facts from text using a predefined KB schema. In that case, both entities and relations must align with valid schema identifiers. This paradigm originates in early automatic content extraction efforts, which established the foundational link

between schema-driven extraction and knowledge-base construction (Grishman 1997, Doddington et al. 2004).

Contemporary systems operationalize this paradigm by grounding extracted triples in large-scale KBs such as Wikidata. According to Josifoski et al. (2022), the objective is to extract a comprehensive set of $(subject, relation, object)$ triples that are consistent with the entities and relations defined by the schema. The methodological implications of this approach will be examined in detail in Chapter 5.

In contrast, open IE was developed to address the limitations of schema-based methods, particularly their rigidity and reliance on domain-specific ontologies. As introduced by Etzioni et al. (2008), open IE systems produce surface-level tuples without presupposing a fixed set of relation types. This is particularly fitting for extraction tasks across heterogeneous corpora.

The ObliquER pipeline developed in this dissertation adopts an intermediate methodological stance in which extraction targets are anchored in the Viewsari ontology schema, while the relational vocabulary remains deliberately underspecified to reflect the interpretive and open-ended nature of claims in *The Lives*. We discuss the rationale for this design in Chapter 7.

4.4 From rule-based systems to transformers: an overview

The subtasks outlined above, ranging from preprocessing to relation extraction, are embedded in the technological paradigms that operationalize them. Each has been shaped and, in some cases, redefined by successive shifts in the underlying computational infrastructure. The following section explains the methodological evolution from early rule-based systems to transformer-based architectures that now dominate the field, situating our ObliquER extraction approach developed in this dissertation.

Benchmarks, in particular Message Understanding Conferences (MUC) (Chinchor 1998), CoNLL-2002 (Tjong Kim Sang 2002), and CoNLL-2003 (Tjong Kim Sang and De Meulder 2003), have become standard evaluation frameworks and remain relevant to the field to this day. A subsequent milestone was automatic content extraction (Doddington et al. 2004), which went beyond surface-level entity detection by incorporating coreference resolution and formalized relation extraction.

The introduction of deep learning in the 2010s introduced recurrent and bidirectional long short-term memory architectures, which extended contextual modeling beyond local token windows but still required substantial labeled training data (Lample et al. 2016). The dominant paradigm shifted

decisively with the advent of transformer-based models such as BERT (Devlin et al. 2019). Pre-training on large corpora followed by task-specific fine-tuning yielded significant improvements in NER and relation extraction across domains (Tenney et al. 2019). In DH research, the application of transformer-based models to historical corpora has enabled large-scale entity recognition and metadata enrichment previously unattainable (Schweter and Baiter 2019, Ehrmann et al. 2023). We further elaborate on state-of-the-art applications of these advancements in Chapter 5, This advance has redefined the scope of computational analysis in the field and has shaped our contribution, as established in Chapter 7.

4.5 Large language models and generative extraction (for historical sources)

Chapter 2 established that LLMs are formally autoregressive, probabilistic text generators whose outputs systematically depend on prompt design, model version, and the distribution of training data. This section analyzes how these architectural properties shape the processes of information and KE that underpin the Viewsari pipeline. In particular, it demonstrates that the architecture introduces not only technical opportunities but also epistemic challenges, as the variability inherent in model outputs foregrounds the interpretive dimension of extraction. Addressing these challenges is central to the methodological framework advanced in this dissertation.

Generative IE constitutes a fundamental departure from the modular, sequential pipelines outlined in Section 4.4. Instead of assigning labels to predefined spans, an LLM receives both the input document and a natural-language task instruction. The model then produces a structured output sequence that encodes entities, relations, and their assertions by generating tokens in an autoregressive manner (Xu et al. 2024). This shift reconfigures the extraction process from a sequence of discrete labeling steps to a unified generative procedure, in which entity types that can be described in natural language can be extracted without pre-training on fixed label sets. The prompt itself replaces the traditional definitions of type and label, enabling the model to resolve references that lack explicit surface forms. Rather than matching a span, the model generates a canonical description or identifier inferred from context. This capability marks a critical advance toward implicit entity recognition (Josifoski et al. 2022, Zhang et al. 2025e, Hosseini 2019). The current state of these approaches, also in CH and humanities contexts, is reviewed in Chapter 5.

These two ways of producing an extraction output rest on two complementary aspects of the same mention. A mention's *span* is the portion of the source text being targeted, a contiguous sub-sequence of one or more tokens identified by its character offsets, whereas its *surface form* is the concrete string of characters that occupies that span. In "a panel for the friars of S. Francis",

the artwork mention spans multiple tokens. Essentially, an implicit mention is an *extensive descriptive span without a specific surface form.* A real-world entity is often realized across multiple tokens, which is why the targeted unit is a span rather than a single token (Ding et al. 2025). The two aspects ground two recognition and linking paradigms, separated by what a system targets as the identity of a mention.

The adoption of LLMs entails a structural epistemic shift central to this dissertation's argument. For example, the phrase "the gold-crested angel in the chapel ceiling" refers to a specific artwork without a canonical title. A surface form-level classifier fails to identify this mention because it lacks training signals. By contrast, a generative model, when prompted with the Viewsari ontology and definitions for classes such as `implicit artwork mention`, can infer that the phrase denotes an artwork and generate a structured representation based on the given context. So, given that in generative models, extraction is inherently probabilistic, outputs depend on factors such as prompt design, model version, temperature, and training distribution, none of which are intrinsic to the source text. Two differently prompted runs of the same model on the same passage may yield divergent entity hypotheses. This variability does not render generative extraction unreliable; rather, it establishes extraction as an interpretive act that requires explicit documentation. Consequently, provenance must encompass not only the source text but also the computational agent and its configuration. This requirement has been necessary before, but it is now more urgent in the context of generative models. We take this as a foundational design principle of the Viewsari ontology, as formalized in Chapter 7.

4.6 Beyond the closed-world assumption: long-tail and out-of-knowledge-base entities

A major challenge in applying extraction tasks to historical sources is the insufficient documentation or canonization of the knowledge embedded in these texts. These entities represent the long tail of CH knowledge (Boscariol et al. 2025, Orgel et al. 2015). This issue is illustrated by the skewed historical representation in *The Lives*, where only a few well-documented Italian Renaissance figures, such as Michelangelo, Raphael, and Leonardo da Vinci, appear frequently and can be matched to established KBs. In contrast, most entities are rare, lack canonical identifiers, and receive little or no coverage in structured repositories (Santini et al. 2022, Ondraszek 2025).

The term 'long-tail' refers to frequency distributions in which the combined volume of low-frequency items greatly exceeds that of high-frequency items (Garigliotti et al. 2019, Chen et al. 2023). In the context of CH KE, the long tail refers to the substantial population of entities that are

individually rare within a corpus but collectively constitute the majority of its content (Boscariol et al. 2025, Graciotti et al. 2025).

Conventional EL systems presuppose that detected mentions can be mapped to stable identifiers within a fixed reference KB. In the case of The Lives, however, such identifiers are absent for most entities. Systems optimized for high-frequency entities that can be resolved against established identifiers typically do not mitigate this bias (Agarwal et al. 2022, Parisot et al. 2022, Graciotti et al. 2025). Consequently, well-known entities such as the 'Mona Lisa' are reliably recognized, whereas less frequent items, such as the 'panel for the friars of Santa Croce, remain unlinked and effectively invisible within the knowledge representation. Accordingly, KG representations must accommodate entities that lack external authority identifiers, allowing them to exist provisionally on the basis of contextual evidence. This requirement foregrounds the need to model entities as interpretive constructs rather than as fixed referents anchored exclusively in external databases.

This challenge has been the subject of sustained investigation in the field of KG completion, with the focus shifting toward methods that can represent and reason about entities not present in the reference KB. Hamaguchi et al. (2018) formalize the challenge of embedding unseen entities without retraining the entire model. Their approach uses graph neural networks to propagate information from known neighboring entities, thereby facilitating the construction of inductive embeddings at test time. Similarly, Amador-Domínguez et al. (2021) propose ontology-based initialization with semantic type information to generate embeddings for unseen entities. More on this as baseline for the ObliquER extraction pipeline and the resulting Viewsari KG is established in Chapter 5 and Chapter 7.

4.7 Evaluation of extraction systems for domain-specific and historical sources

Benchmark corpora have fundamentally structured the evaluation of extraction systems, as discussed in Section 4.4. MUC introduced task-oriented evaluation protocols and automatic scoring mechanisms for both entity and template extraction (Chinchor 1998). Building on this foundation, subsequent frameworks such as CoNLL-2003 for NER (Tjong Kim Sang and De Meulder 2003) expanded the evaluative scope, thereby establishing a trajectory toward increasingly comprehensive assessment standards. The introduction of multi-layer corpora such as OntoNotes (Pradhan et al. 2007) facilitated more comprehensive evaluation regimes, notably incorporating coreference resolution as exemplified by CoNLL-2012 (Pradhan et al. 2012). As KE becomes central, evaluation criteria must move beyond this correctness to assess quality, consistency, and completeness in the resulting knowledge structures.

Later approaches shifted the focus toward EL and KB population (Ji and Grishman 2011), while SemEval tasks provided fine-grained benchmarks for semantic relation classification and other subtasks (Hendrickx et al. 2010). Large KGs such as DBpedia (Auer et al. 2007) and Wikidata (Vrandečić and Krötzsch 2014) contribute to evaluation by providing data for large-scale assessment of alignment quality, linking accuracy, and the coverage of extracted triples against reference KBs.

Most tasks use evaluation metrics such as precision, recall, and F_1 score, which are usually reported as micro- and macro-averages. In NER, strict span matching is often compared with partial- or overlap-based scoring. In EL, accuracy and clustering-based measures are common (Tjong Kim Sang and De Meulder 2003, Pradhan et al. 2012).

In DH and CH contexts, however, these standard benchmarks are often insufficient, and the evaluation metrics are not sufficiently flexible for the nature of textual sources. A detailed review of this will follow in Chapter 5.

5 State of the art in generative knowledge extraction and knowledge engineering

This chapter surveys state-of-the-art approaches in information and KE, semantic web technologies, and DH research, and compares them to the approaches of this dissertation. While Chapter 2–Chapter 4 introduced the theoretical and technical foundations, this chapter focuses on contemporary approaches (ca. 2024–2026) and their methodological implications. The structural mismatch between contemporary IE and KG construction pipelines, which are optimized for explicit extraction, forms the central problem space.

5.1 Generative knowledge extraction

Generative knowledge extraction reconceptualizes IE as a conditional generation task. This section begins with a survey of LLM families that underpin contemporary extraction systems. It then examines the prompting and decoding strategies used to operationalize these models, with particular attention to the requirements of interpretive historical sources.

5.1.1 Overview of contemporary large language model families

The literature surveyed in this chapter is based on a heterogeneous set of (large) language models. In this section, we define the differences among existing architectures and model families based on accessibility (open-weight versus proprietary) and training generality (general-purpose versus domain-adapted).

The most widely used open-weight models include Meta's LLaMA family (Touvron et al. 2023, Grattafiori et al. 2024), Mistral AI's Mistral and Mixtral series (Jiang et al. 2023), and Alibaba's Qwen family (Qwen Team et al. 2025). These models make their parameter sets publicly available,

enabling local deployment, fine-tuning, and auditing of model behavior. This openness supports the reproducibility and transparency that are important for scholarly work. For these reasons, this dissertation uses LLaMA to extract implicit entities, as described in Chapter 7.

Proprietary models, including those accessible via OpenAI's API (GPT-3.5, GPT-4, GPT-4o, etc.) (OpenAI et al. 2023), Anthropic's Claude series (Anthropic 2024), and Google DeepMind's Gemini family (Gemini Team, Google et al. 2024), do not release weights. As a result, their outputs cannot be reproduced. Even the same prompt at different times might yield different results. The training data distributions are also not shared.

The distinctions between proprietary and open models, as well as their respective performance, are critical for humanities extraction pipelines. For instance, generative extraction of implicit artwork references or the resolution of descriptive phrases about artworks to Wikidata (Vrandečić and Krötzsch 2014) entities depends not only on the prompt and input text but also on the model's latent training data. In proprietary models, the contents of the training corpus are often unknown, which increases the risk of training data contamination. A model may appear to infer information from the input, but it may instead extract information it has memorized (Kandpal et al. 2023). This issue is especially pertinent for widely available sources such as Giorgio Vasari's texts, which have been digitized and translated for decades. Such sources can distort information and produce biased outputs, such as the frequent attribution of artworks to Michelangelo rather than to lesser-known artists.

Encoder-only architectures, including BERT (Devlin et al. 2019) and its variants such as CamemBERT (Martin et al. 2020) and GHisBERT (Beck and Köllner 2023), are standard baselines for NER and EL tasks in IE.

In contrast, large-scale decoder-only and encoder-decoder models, such as the aforementioned GPT (OpenAI et al. 2023) or Qwen (Qwen Team et al. 2025), have become prevalent in zero-shot and few-shot extraction scenarios, reflecting a shift toward generative IE without extensive task-specific supervision (Pan et al. 2024, Kandpal et al. 2023).

Generally accessed through APIs, these models are not subject to further fine-tuning unless explicitly indicated, as with distilled models such as UniversalNER (Zhou et al. 2024c). Within the Viewsari pipeline, and specifically in ObliquER (Chapter 7), the focus is on prompt-based extraction using instruction-tuned models. The openai-gpt-oss-120b model was selected as the principal model based on its demonstrated ability to achieve robust zero- and few-shot performance on complex generative information-extraction tasks (OpenAI 2025). For the candidate retrieval stage in EL, Qwen/Qwen3-Embedding-4B was adopted due to its efficiency and empirically validated effectiveness on historical text embeddings (Zhang et al. 2025b).

5.1.2 State of the art of generative information extraction

In classical IE systems, labeling spans or classifying relations between pre-labeled entities is performed discriminatively based on the input sequence. Within this paradigm, extraction is fundamentally a process of detection: the model identifies elements already instantiated in the text according to a predetermined typology (Nadeau and Sekine 2007). This framing presupposes that the entities and relations of interest are surface-realized, as explicit mentions, which is precisely the condition that breaks down in the extraction setting addressed by this dissertation, where entities are often referenced through implicit mentions.

Generative IE, defined in Chapter 4, departs from this architecture: rather than labeling pre-given spans, the model generates structured output conditioned on an input document and a natural-language instruction, with the schema supplied as a soft specification through the prompt rather than as a hard architectural constraint (Josifoski et al. 2022, Han et al. 2024, Zhang et al. 2025e). The remainder of this section surveys this shift and its consequences for the modeling choices adopted in ObliquER (Chapter 7).

This approach enables open-ended extraction across entity types and relation categories not anticipated during training and allows the model to handle even implicit or descriptive references that normally resist span-level detection (Liu et al. 2024). This distinction is significant for the present dissertation in two respects: first, the ObliquER pipeline operates in the generative paradigm, as our system employs an LLM to identify implicit entity references by conditioning generation on a prompt derived from the Viewsari ontology schema.

Second, we underline that the shift from discriminative to generative extraction alters the epistemic status of extracted outputs, because classifiers assign deterministic labels to spans, but generative models produce probabilistic continuations shaped by prompt design, training distribution, and model configuration, which are precisely the conditions that must be documented in a provenance-aware KG (Section 2.4).

As established in Chapter 4, LLMs have become versatile replacements for task-specific supervised architectures, including sequence-labeling models for NER. They can be flexibly adapted to new domains or document types by providing an updated schema or a minimal set of demonstration examples in the prompts, which enables substantially faster deployment in real-world settings, as demonstrated in clinical NER (Hu et al. 2024).

Zhang et al. (2025e) and Xu et al. (2024) have surveyed recent developments in standard IE subtasks within the generative framework. They document the rapid adoption of this paradigm across all subtasks and summarize methodological advancements toward settings where structured targets are converted into textual sequences via autoregressive sequence generation. For instance,

consider the sentence “Giorgio Vasari was born in Florence.” A traditional system might use BIO tagging to label the tokens as follows: [Giorgio/B-PER, Vasari/I-PER, was/O, born/O, in/O, Florence/B-LOC, ./O]. In contrast, a generative model could output a JSON-style string: {“Person”: “Giorgio Vasari”, “Place_of_Birth”: “Florence”}. Through more flexible input and output transformations, it has become possible to unify diverse tasks under a single generative objective, facilitating multi-task or unified information-extraction models without requiring an architectural redesign for each new schema (Liao et al. 2025).

Zhang et al. (2025e) also emphasize that advancements in these models can be traced back to model scaling, which plays a critical role, as noted by Kaplan et al. (2020) and revisited by Chen et al. (2025b). Empirical evidence indicates that when model sizes exceed a specific parameter threshold, in-context learning performance on information-extraction tasks improves significantly. Larger models exhibit enhanced adaptation and cross-domain generalization, often enabling zero-shot or few-shot extraction without explicit fine-tuning. This scaling phenomenon has contributed to the increasing prevalence of prompt-based extraction methods, in which task descriptions, schemas, and output formats are specified directly via natural-language instructions (Han et al. 2024).

While Zhang et al. (2025e) demonstrate that fine-tuned pre-trained language models remain competitive with, and often surpass, much larger language models on in-domain IE benchmarks, LLMs possess a clear advantage in generalizing to unseen schemas, unfamiliar domains, and low-resource settings, as shown by zero-shot transfer results on datasets like MIT Movie,[1] which is the setting that we require for ObliquER, whose primary challenge is the extraction of implicit entities from Vasari’s *The Lives* in a low-resource setting. No annotated corpus currently exists for implicit entity detection in this domain, and the manual annotation effort required to construct one at the necessary scale for model fine-tuning would be prohibitive. Each implicit mention must be resolved within the context of Italian Renaissance art, a process that demands slow, expert human annotation. The GT corpus developed for evaluation in this dissertation represents a substantial annotation effort, yet it remains orders of magnitude smaller than the training sets used by competitive fine-tuned IE models. Under these constraints, such references are inaccessible to fine-tuned, pre-trained models, which are typically trained on surface-level named entity corpora. Resolving implicit entities requires broad contextual knowledge that only large-scale pretraining provides. Furthermore, the bespoke Viewsari schema precludes reliance on any existing annotated resource. Analogous low-resource extraction challenges arise wherever bespoke schemas meet specialized source material in the absence of large annotated corpora, and the methodological responses mirror those adopted in our approach. Gesese et al. (2026) study how to extract research artifacts from GitHub README files using a custom ten-type schema

[1] MIT Movie dataset on Huggingface: `https://huggingface.co/datasets/tner/mit_movie_trivia`.

based on an ontology. They annotated 200 documents in INCEpTION (Klie et al. 2018). Their results show that fine-tuned transformers outperform zero-shot LLMs on domain-specific entity types, with exact-match F1 scores rising from 0.23 to 0.72. This suggests that even a small amount of annotation can help restore the benefits of task-specific training, especially when schemas differ markedly from those seen during pretraining. ObliquER uses fewer annotations and a schema that requires more interpretation, making in-context prompting for large models more useful. Still, the overall pattern in a low-resource setting evidently appears in many different domains. For *The Lives* specifically, Santini et al. (2022) have tested a distillation-based approach using UniversalNER (Zhou et al. 2024c) for artwork and subject recognition, to which we return in Section 5.1.5 and Section 5.2.4.

Generative IE offers increased flexibility and schema adaptivity, yet it still presupposes convergence toward a single canonical structured representation. This assumption fails to accommodate interpretive multiplicity and epistemic stratification, both of which are foundational in humanities corpora. As a result, methodological challenges arise when structured targets are fundamentally unordered sets, and merging them into sequences introduces order bias and amplifies decoding errors. Winkle et al. (2026) show this in their adoption of generative extraction for German biographical encyclopedia entries that abandons strict surface form matching for embedding-based comparison between extracted and gold values, on the grounds that exact-match scoring penalizes the systems for legitimate paraphrastic variation (“Children’s books” vs. “Children’s and young adult literature”) and treats annotation gaps as system errors. Their findings are instructive in two further respects relevant to ObliquER. First, prompt engineering strategies offered limited gains in setup and formulation; clear formulation and few-shot prompting yielded the most consistent improvements, while chain-of-thought (CoT) prompting and self-refinement provided little benefit, and multi-step decomposition actively degraded performance. Second, manual error analysis revealed that many apparent false positives reflected missing gold annotations rather than factual errors, a pattern that any evaluation regime for low-resource humanities IE must contend with. Both findings reinforce the case for treating extraction outputs as provenance-bearing artifacts whose evaluation cannot rest on token-level matching alone.

Evaluation practices remain tethered to token-level or triple-level exact match metrics, which were developed for deterministic classifiers and do not adequately reflect the stochastic, distributional nature of generative outputs (Jiang et al. 2024, Fan et al. 2024, Zhang et al. 2025e). There is, therefore, a growing recognition that generative IE entails not only a shift in model architecture but also a reconceptualization of what constitutes an extraction output and how it should be evaluated. Persistent challenges in generative IE include maintaining output consistency and ensuring compliance with structured format specifications (Tam et al. 2024).

The work surveyed in this section reframes extraction from span detection to conditional generation, but the entity-as-target remains the same object as in the discriminative paradigm. The schema is supplied as a soft prompt rather than a hard architectural constraint (Josifoski et al. 2022, Han et al. 2024), but the extracted entity is still treated as a referent already present in the source whose recovery is the model's task. Generative IE has, in this sense, inherited the assumption it could most easily have unsettled: the *pre-interpretive entity assumption* (**C1**); the surveyed approaches (Zhang et al. 2025e, Xu et al. 2024, Han et al. 2024) differ on the model, the prompting strategy, and the schema, but converge on the assumption that the mention-to-referent resolution is the model's deliverable rather than the representational work being done. ObliquER (Chapter 7) departs from this convergence by treating the resolution itself as a documented interpretive activity (**C2**), grounded in the three-layer architecture of the Viewsari ontology.

A parallel response to the hallucination problem in generative IE has emerged from anchor-constrained, agentic decomposition. Yang et al. (2026) introduce the Anchor–Extraction–Verification–Supplement (AEVS) framework, in which extraction is staged across three LLM calls: anchor discovery identifies entities, relation phrases, and attribute values with their positions in the source text; grounded extraction constrains triplet generation to those discovered anchors; and a restoration-based verification step, paired with a coverage-aware supplement, validates the output against the input. Across WebNLG, REBEL, and Wiki-NRE, AEVS reports per-configuration hallucination rates between 0.23% and 20.23% at a cost of 2.83–4.28 LLM calls per sample, and ablations confirm that the anchor constraint, rather than the verification step, is the principal mechanism for hallucination reduction. The AEVS framing, that "the root cause [of hallucination] is the lack of provenance: existing methods produce triplets without explicit links to their textual origins" (Yang et al. 2026), aligns directly with the diagnosis advanced in this dissertation, and the staged, agentic decomposition exemplifies the broader trend toward tool-augmented and multi-step extraction pipelines surveyed by Plaat et al. (2025), Zhang et al. (2025a), Liu and Shu (2026). The two systems diverge, however, on the kind of provenance recorded. AEVS treats anchors as token-position pointers into the source and validates triplets against an assumed referent already extractable from the surface text, a regime made coherent by the explicit-fact character of its evaluation corpora. ObliquER, in contrast, must record *interpretive* provenance – prompt template, model version, mention typology (explicit, implicit, coreferent, generic), and ontology layer under which a descriptive phrase is identified as an artwork mention – because for the implicit references that dominate *The Lives*, the surface span is not the referent, and anchor-constrained generation cannot resolve what is descriptively present but referentially absent. AEVS is, in this sense, a confirmation of the provenance turn that motivates the present pipeline (**C2**) and an illustration of its limits when the source is interpretively dense rather than factually explicit; the anchor constraint is the PIEA (**C1**) operationalized at the token level.

Another important issue, which is mostly missing from the literature, is transparency. Outputs from generative extraction pipelines often end up in datasets without any record of the model, prompt, decoding settings, or source span used to create them. This makes it impossible to trace how a specific assertion was made. Some models, such as PROV-O by Lebo et al. (2013), have already addressed this. ObliquER extends this approach by treating each extracted statement as a provenance-bearing artifact, built on the ontology for that extraction. The prompting strategies described in the next section are thus not just a way to get structured output. They are also a key part of the provenance record and the interpretive process the model uses to identify entities from implicit mentions.

5.1.3 Prompt engineering for generative information extraction

Prompting is key to helping language models produce structured output from unstructured text, which is crucial for high-quality IE. While Chen et al. (2025a) systematically analyze prompt engineering strategies for LLMs and shows how structured prompts, schema constraints, and temperature settings affect the output, IE poses additional challenges. Here, the model must resolve ambiguities in the source text and adhere to strict output formats, such as JSON for entities or knowledge triples (Tam et al. 2024, Zhao et al. 2025a). This section examines prompting methods, focusing on their effectiveness for entity recognition, relation extraction, and knowledge triple generation.

Polat et al. (2025) study prompt engineering for text extraction by testing five prompting methods with seventeen different prompt templates. They evaluate these on three LLMs (GPT-4, Mistral 7B, Llama 3) using the RED-FM relation extraction dataset (Huguet-Cabot et al. 2023). Their results, when checked against Wikidata's data schema, offer the most thorough comparison of prompting strategies for KE to date. Zhao et al. (2025a) add to this by introducing Active Prompting for Information Extraction, a framework that looks at how the choice of examples in prompts affects results. Together, these studies ground this section in the specific needs of extraction tasks, not just the general language generation of language models

The following subsections dive deeper into the strategies that can be used to perform extraction tasks with generative models, e.g., zero-shot prompting in Section 5.1.3.1 or advanced techniques in Section 5.1.3.3.

5.1.3.1 Zero-shot prompting for extraction

Zero-shot prompting is a paradigm in which a prompt fed to the model provides no examples or demonstrations (Schulhoff et al. 2024); it is an inference strategy or a task instruction in which the model performs a new task solely based on natural language instructions (Brown et al. 2020, Kojima et al. 2022, Fu et al. 2023, Polat et al. 2025).

```
Prompt:
Extract all artwork titles from the following text input.

Text input:
<<In the chapel hangs the Last Supper
and a depiction of the Madonna and Child.>>
```

Cocchieri et al. (2025) present ZeroNER, a zero-shot NER method that uses natural language descriptions of entity types to a frozen LLM instead of performing task-specific fine-tuning. This approach allows the system to adapt to new entity categories by using type descriptions for extraction. Polat et al. (2025) create a zero-shot baseline with a direct instruction prompt ("Extract knowledge triples from the text. Return the triples in JSON format.") and show that it provides a strong lower bound, especially with advanced models like GPT-4. Both findings are surprisingly robust; thus, zero-shot prompting as a technique is well-suited to settings where no annotated training corpus of sufficient scale exists to support the alternative, which is particularly useful in low-resource settings like ObliquER and Vasari's *The Lives* (Polat et al. 2025).

5.1.3.2 Few-shot prompting for extraction

In few-shot prompting, the prompt includes one or more input <–> output examples before the actual task query, which, as in-context learning signals, guide the model toward a desired output structure without updating model parameters. Instead, it conditions the generation process by inducing contextual patterns (Brown et al. 2020, Schulhoff et al. 2024).

The effectiveness of few-shot prompting for extraction tasks depends on the selected examples. Polat et al. (2025) demonstrate that incorporating retrieved examples selected via maximal marginal relevance from the training data significantly enhances extraction performance across all tested models. This has also influenced our design decision for prompting for implicit entities with ObliquER, as we explain in Chapter 7. Additionally, they show that few-shot prompts enhanced with retrieval-augmented generation consistently outperform both canonical few-shot and more complex reasoning-oriented methods, such as CoT prompting, suggesting that examples have the

greatest impact on output quality. This finding is validated by Zhao et al. (2025a), underscoring that few-shot extraction is highly sensitive to example selection and that common strategies, such as random sampling, often prove suboptimal due to the structural and semantic complexities inherent in extraction tasks (Liu et al. 2021).

Interestingly, empirical studies have shown that zero-shot prompting can, in some cases, outperform few-shot prompting, particularly when examples introduce bias or constrain model generalization (Chen et al. 2025a).

As briefly mentioned above, in Chapter 7, we elaborate on the choice of prompting techniques for the ObliquER extraction pipeline, focusing primarily on few-shot prompting with maximal marginal relevance for selecting training data from the GT to achieve the best results. However, we also apply more sophisticated prompting styles, especially since provenance plays such a crucial role in the resulting KG. For this, CoT prompting, in which the model explains its reasoning, is particularly relevant for foregrounding the interpretive process.

5.1.3.3 Advanced prompting modes

Prompt engineering has moved beyond zero- and few-shot instructions, with approaches that enable the detailed surfacing of the model's reasoning processes, helping to control outputs more effectively. These advancements have been motivated primarily by the fact that training data is often opaque for LLMs, and distillation on it is costly, so control via prompting is cost-effective (Liu et al. 2023a, Schulhoff et al. 2024, Zhang et al. 2025e, Gu et al. 2026). Among these advancements are structured reasoning techniques such as CoT prompting (Wei et al. 2023), self-consistency (Wang et al. 2023b), and role prompting (Shanahan et al. 2023), as well as automated prompt optimization frameworks (Zhou et al. 2023), of which we will focus on CoT prompting and automated prompt optimization specifically. In the following subsections, we examine the functionalities and improvements introduced by these approaches and connect them to the prompting strategies used in ObliquER.

Chain-of-thought prompting CoT prompting generates intermediate reasoning steps before producing a final answer. This process can be initiated with explicit scaffolding phrases, such as "Let us think step by step," which have been shown to improve performance on logical and arithmetic reasoning tasks (Schulhoff et al. 2024, Polat et al. 2025, Chen et al. 2025a).

Polat et al. (2025) adapt this approach specifically for extraction tasks by using triple extraction across three intermediate steps, namely entity extraction, entity typing, and relation extraction, mirroring the traditional IE pipeline in one single prompt. CoT prompting can be done in different

prompting settings, such as zero- or few-shot. In a zero-shot setting, they instruct the model with "Before extracting triples, let's think step by step", which guides it towards sequential reasoning without examples from the application text. Similarly, in a few-shot setting, the retrieved examples can be fed back to the model to illustrate how intermediate steps must be executed.

The empirical record on CoT prompting for extraction tasks is mixed. Polat et al. (2025) do not observe a reliable outperformance of simpler instruction-based prompts with retrieved examples, suggesting that explicit reasoning may not be necessary when high-quality examples are available. Winkle et al. (2026), evaluating CoT prompting on generative IE from German biographical encyclopedia entries, report only marginal F1 improvements over a baseline configuration, with gains in recall offset by losses in precision; few-shot prompting and example count proved to be the dominant performance drivers in their setup, while CoT reasoning and self-refinement contributed comparatively little.

In the context of this dissertation, however, CoT prompting is adopted not to improve extraction scores, but for reasons that the existing literature has not foregrounded. First, implicit entity recognition is inherently an *interpretation*: identifying "the minister of culture to the Medici court in Florence" as a reference to Giorgio Vasari requires reasoning over context and historical background knowledge, none of which a span-level detector can access. By eliciting an intermediate reasoning trace such as "the phrase *the minister of culture to the Medici court in Florence* is used elsewhere in this biography to refer to Vasari; therefore, this instance likely refers to the same individual," a CoT prompt makes this inferential chain explicit (Kojima et al. 2022, Wei et al. 2023).

Second, and more consequentially for the design of ObliquER, CoT outputs furnish provenance for the interpretive steps taken by the language model. If the reasoning trace is retained alongside the extracted entity, the resulting KG records not only what was extracted and by whom, but also the basis on which the extraction was made. This reframes the reasoning trace from a prompt-engineering technique aimed at performance into a constitutive provenance object, directly supporting the transparency requirements set out in the Viewsari ontology (Chapter 7). The mixed extraction-score results reported by Polat et al. (2025) and Winkle et al. (2026) are therefore not in tension with our use of CoT prompting: where they evaluate the technique against task accuracy, we evaluate it against interpretive transparency, and the trace is valuable to ObliquER even in cases where it does not improve recall or precision.

Role prompting. In addition to the instruction concerning what to extract, role prompting assigns a contextual identity to the model, such as historian, legal scholar, or biomedical expert (Schulhoff et al. 2024).

Automated prompt optimization Recent methods aim to refine prompts automatically using strategies such as Prompt Optimization with Textual Gradients (ProTeGi) (Pryzant et al. 2023) or reinforcement-learning-based refinement (Chen et al. 2025a).

ProTeGi, Pryzant et al. (2023) propose a prompt-refinement framework to overcome reliance on manual trial-and-error design, which typically requires humans to verify outputs from LLMs and then adapt the prompts to yield better results. This reframes prompt design as a structured optimization problem, inspired by numerical gradient descent, in which the gradients simulate natural-language explanations of why the current prompt fails on training examples.

For instance, if a prompt for artwork extraction fails to detect implicit descriptions, a textual gradient might state:

> "The prompt assumes artworks are always explicitly titled, whereas the text contains descriptive references lacking canonical names."

Afterward, the model is instructed to edit the prompt in the opposite direction of the gradient's semantics. In other words, it revises the prompt to address the flaw or issue identified by the textual gradient. If the gradient detects that implicit detection is missing, the revised prompt can instruct the model to consider descriptive passages as candidate artworks (Pryzant et al. 2023).

Uncertainty-aware prompting for information extraction The studies reviewed above treat prompting strategies as fixed design choices, varying only in method or exemplar retrieval mechanism, without fully inquiring into the model's uncertainty about a given input. Zhao et al. (2025a) argue that this static view overlooks a fundamental challenge for models to stick to rigid output schemas.

To address this gap, Active Prompting for Information Extraction leverages the concept of introspective confusion. Here, the model is prompted to reflect on its own generative process, thereby revealing the inputs it is most uncertain about. This approach highlights two types of uncertainty in generative extraction: format-level uncertainty, which concerns producing structurally correct outputs, and content-level uncertainty, which examines how consistent the extracted information is across different generation attempts using set-based Jaccard divergence. By measuring these uncertainty scores, the method ranks unlabeled samples by difficulty. The most difficult examples are then selected as few-shot examples and incorporated into a new prompt template. This approach demonstrates superior performance over other baselines, e.g., zero-shot, on four IE benchmarks and LLMs (Zhao et al. 2025a).

Generative information extraction as annotations Generative IE can also be understood as an interpretive annotation task. Traditionally, annotation has been performed by humans, whose work serves as the GT for model evaluation. When LLMs undertake this task, they similarly map text to predefined categories, and it is necessary to model these interpretive actions in a comparable manner. Recent research demonstrates that LLMs can serve as effective annotators. Studies indicate that these models frequently align with human annotators on span-level tasks and, in some cases, match or exceed the performance of crowd workers (Kasner et al. 2026). In the context of clinical IE, increasing the number of samples annotated by LLMs has improved downstream model performance and surpassed manual annotation on three of four evaluated tasks (Shimizu et al. 2025).

Prompt strategies as provenance objects. The prompt-engineering literature surveyed in this section evaluates strategies against extraction accuracy: which prompt yields the highest F1, which exemplar selection most reduces variance, which structural template best constrains output (Polat et al. 2025, Zhao et al. 2025a, Winkle et al. 2026). This is the right evaluation regime for a fact-recovery task, but it is the wrong regime for the task ObliquER addresses. Where extraction is itself an interpretive act, prompts, exemplars, and CoT traces are not only inputs to a performance score but also constituents of the interpretive record: they document the conditions under which a particular reading of an implicit reference was generated. We refer to this representational deficit, the fact that current practice produces interpretations whose interpretive conditions are not retained, as the *provenance gap* (**C2**).

5.1.4 Going obliquer: Existing approaches for implicit entity recognition

Implicit communication is a key part of how people interact. Pragmatics and cognitive linguistics show that meaning comes from context, background knowledge, and social interpretation, not just from the words themselves (Fischer 2017).

Implicit entity recognition, as defined in Chapter 2 and described in the ObliquER task in Chapter 7, means finding entity mentions that do not have a standard surface form and need context to be understood. In this section, we review current approaches.

Extracting implicit knowledge from text, known as implicit IE, has long been a major, unresolved problem in NLP. Early methods suggested recovery of implicit content through additional domain context (Cavazza and Zweigenbaum 1992). These systems relied on incorporating KBs to infer data, since handling lexical surface forms was not enough.

While standard approaches in contemporary IE continue to evolve and improve, especially compared with early approaches, the solution to this problem remains as oblique as the mentions themselves. To this day, despite a rediscovery of the problem as central to IE, especially in generative systems, performance remains low and unstable (Stramiglio et al. 2025). Thus, most systems continue to assume stable entities with surface forms detectable via span identification, followed by candidate ranking against a KB. This assumption fails critically like rolling a Critical 1 in Dungeons and Dragons when texts include descriptive references or OOKB entities (Hamaguchi et al. 2018).

Of particular interest for solving this problem are approaches to open-world and fine-grained entity recognition, where entities are inferred from context, thereby reframing the task as a natural-language inference problem (Schimmenti et al. 2025). Nevertheless, these approaches still presuppose identifiable mention spans without regard for the interpretive processes underlying an LLM's determination that a description is an entity reference and, if so, the fitting entity.

Some of the earliest formulation of implicit entity recognition addressed clinical documents (Perera et al. 2015), framing the task as sentence-level classification against a fixed set of target entities whose models are built from KB definitions rather than from running text, and explicitly treating boundary detection as an optional, deferred step. It shares the core intuition that a referent can be evoked without its name, yet it operates over a closed target inventory, resolves no OOKB referents, and records none of the extraction conditions under which a reading was produced. Notably, that work observes that no baseline could be applied for a fair comparison of this novel task, which is the same condition under which the surface form-level systems of Section 9.4.2 serve as the closest available comparators rather than competing implicit recognizers.

Recent work has identified a widening gap between generated output, surface form, and source text in LLMs, attributable to implicit communication. As Tint et al. (2026) demonstrates, these models encode and decode information implicitly, shifting the focus from literal string matching to interpretive processes. As established in Chapter 2, this requires treating implicit entity recognition not as technical span detection but as an epistemic act. What remains open, however, is a formal representation of these inferential processes – the gap this dissertation addresses through provenance-aware modeling.

Stramiglio et al. (2025) found that, based on a controlled evaluation using Wikidata-derived biographical texts, LLMs exhibit measurably reduced performance when required to extract information from implicit formulations. Their results underscore that the implicitness of a source text is not a neutral stylistic feature but a structural property that conditions the reliability of extracted outputs, which is precisely the concern motivating the provenance-aware treatment of implicit entity resolution in ObliquER. This absence poses a significant challenge for humanities scholarship, where entities are often introduced through divergent narratives, and their provenance

cannot be reconstructed without interpretive intervention. Implicit entity recognition, therefore, must be understood not as a mere technical extension of NER, but as a modeling problem situated at the intersection of IE and epistemology.

5.1.5 Limitations of generative information extraction and traditional metrics for their evaluation

LLMs continue to show remarkable few-shot capabilities. However, recent empirical studies, e.g., Han et al. (2024), show that these capabilities remain limited when applied to structurally complex information-extraction tasks. Essentially, this stems in part from the fact that the performance of generative systems depends on various factors, including pretrained knowledge and task-instruction design. While recent work has primarily focused on aspects such as model generalization and instruction tuning (Brown et al. 2020, Ouyang et al. 2022), comparatively, only few approaches have paid attention to epistemic status or interpretation in the extracted outputs of LLMs (Mugleston et al. 2025, Liu et al. 2025), and many attempts focus solely on legal aspects and training data provenance (Stober 2025, Liu et al. 2026)

5.1.5.1 Generative information extraction in complex extraction tasks

The study by Han et al. (2024) is divided into 14 information-extraction subtasks, including named-entity recognition, relation extraction, event extraction, and aspect-based sentiment analysis. The results show competitive performance on simpler classification-style subtasks, such as aspect-level sentiment classification, but performance drops for structured prediction tasks, such as relational triple extraction. Importantly, even few-shot prompting and CoT strategies fail to consistently close this gap, as we have emphasized in Section 5.1.3.3.

As Zhao et al. (2025a) observe, the persistent performance gap in generative IE can be traced to a fundamental misalignment between the autoregressive nature of LLM outputs and the rigid structural requirements of formats such as JSON or XML. Traditional extraction systems impose explicit schema constraints and fixed label spaces, ensuring that outputs conform to predefined structures. In contrast, generative models operate without such enforcement, producing outputs that frequently exhibit boundary overextension and partial correctness. Soft-matching evaluation reveals that strict span-based metrics systematically underestimate semantic accuracy, yet even under relaxed criteria, the discrepancy remains. Wang et al. (2025a) extend this analysis by introducing a multi-problem evaluation framework. Their results indicate that while homogeneous classification tasks suffer only minor degradation, performance declines sharply when models must coordinate classification with index selection or structured output generation. This degradation

is not mitigated by prompt engineering strategies such as zero-shot, few-shot, or CoT prompting. The evidence thus points to an intrinsic limitation: generative models struggle to integrate discrete structural operations within the generative decoding process, a challenge that is especially acute in tasks requiring simultaneous span identification, type assignment, and contextual coordination across multiple elements. The study by Winkle et al. (2026) compares LLMs with supervised baselines and finds that generative models often underperform on nuanced relation-extraction tasks compared with smaller, task-specific supervised systems. These results suggest that while LLMs are powerful for isolated extraction tasks, they may struggle with complex structural reasoning across extended narrative contexts.

Task-specific or distilled models address these limitations by forgoing the broad generalization of LLMs in favor of producing more reliable structured outputs within a defined extraction scope. Instead of prompting a general-purpose model during inference, this approach trains a smaller student model on outputs from a larger teacher model, using instruction tuning targeted to a specific application class, such as NER or open IE (Zhou et al. 2024c).

5.1.5.2 Evaluation frameworks for generative information extraction

However, it is not only the extraction tasks themselves that still pose hurdles to generative systems. Given the distinct nature of LLM outputs and the generation process itself, evaluation frameworks used for traditional IE tasks often fail to evaluate newer systems' outputs.

Generative systems approach extraction as sequence generation, generating entities, relations, and triples in text form (Josifoski et al. 2022, Zhang et al. 2025e). According to Wang et al. (2025a), this can lead to outputs that do not match expected forms, often because of boundary overextension or partial correctness. Different prompt setups can also yield markedly different extraction results, even when the system's underlying understanding remains the same (Liu et al. 2021). Since traditional evaluation methods for information and KE were designed for extractive systems that use predefined label sets, metrics such as precision, recall, and F_1 rely on matching predicted spans to gold annotations. However, in this approach, the assumption is based on a given list of expected outputs in a benchmark, and correctness is defined as an exact or near-exact structural match. Naturally, based on what we have established about generative systems so far, this does not match outputs with fuzzy boundaries, so accordingly, strict span-level precision and recall often fail to measure *semantic*, unstable correctness. Small differences in wording can lead to unfair penalties even when the meaning is the same, and fluent but incorrect content may seem accurate at first (Fan et al. 2024, Xu et al. 2024, Kim et al. 2024, Stramiglio et al. 2025).

The magnitude of this underprediction has been quantified in materials-chemistry extraction by Dagdelen et al. (2024), who fine-tuned GPT-3 and Llama-2 for joint NER and relation extraction

across three domain-specific tasks and compared exact-match F1 scores against manual scores assigned by a domain expert. Their general materials extraction task demonstrates the model's ability to normalize entities (*"Li Co O2"* corrected to *"LiCoO2"* without post-processing) and to produce valid paraphrastic variants that the annotation does not record. These are extractions that exact-match metrics classify as failures but that a domain expert would classify as successes. The implication for any humanities corpus is sharper still: where the source text is interpretively rich and annotation schemes must make choices between equivalent surface forms, the penalty imposed by strict matching grows, and the measured performance of a generative system increasingly reflects the annotation's arbitrary choices rather than the extraction's accuracy.

Another limitation arises from benchmark incompleteness. As argued by (Fan et al. 2024), conventional datasets systematically underestimate generative model performance because gold annotations are restricted by predefined schemas and limited annotation budgets. Generative models frequently produce plausible, text-supported triples that do not appear in the gold standard, thereby introducing false negatives under F_1 scoring. This mismatch is essentially one between a closed-world assumption and generative systems that operate in an open-world space. Additionally, in traditional systems, right and wrong are usually clearly labeled, and results are sorted as false positives, false negatives, and similar categories. However, this works differently in LLMs because they operate in unclean water, since they absorb biases from training data and carry them into later tasks, which might not immediately or at all be visible in the results (Li et al. 2024). Instead of probing for mislabeling, the question becomes if a generative system produced an output shaped by patterns it saw during pre-training. Traditional exact-match evaluation cannot address this kind of fuzzy error, which is not a false positive or negative in the usual sense, but a systematically shaped interpretation. This is amplified by the issue that generative models can create hallucinations; they might still receive high scores if they match incomplete or noisy labels, and misleading results can go unnoticed if the evaluation only considers overlap (Huang et al. 2025). Traditional precision and recall do not measure either type.

These changes show that evaluating generative IE and knowledge engineering needs to go beyond mere correctness checks, however, these go beyond the scope of this thesis. Some first steps in this direction include using LLMs for semantic matching and natural language inference to enrich labels (Fan et al. 2024).

5.1.5.3 Neuro-symbolic knowledge extraction – a potential solution?

Generative IE uses LLMs to create structured data from text (Zhang et al. 2025e). These methods, which convert unstructured text into structured formats using neural approaches, often lack clear mechanisms to enforce rules and maintain consistent logic. Neuro-symbolic approaches

address this by adding another layer of control, combining neural language models with symbolic representations based on formal schemas (Hitzler et al. 2024, Yang et al. 2024). In generative IE, the neural models first generate possible answers, and symbolic components check or organize them using an ontology or KB. The results can then be used to guide further steps in the pipeline (Pan et al. 2024).

One example of advanced work using neuro-symbolic methods to improve generative IE is OntoED (Deng et al. 2021). This approach detects event types by using graph neural networks and formal definitions of event classes in an event ontology.

Another area of research evaluates whether LLMs can produce output that matches an ontology, rather than improving extraction with symbolic components during inference. Mihindukulasooriya et al. (2023) introduce Text2KGBench, a benchmark for generating KGs from text using ontologies. In this task, a model receives a set of concepts, standard relations, and domain or range constraints, and must extract triples that match the input sentence, follow the ontology, and avoid hallucinations. The benchmark includes two datasets from Wikidata and DBpedia and measures fact-extraction accuracy using seven metrics: precision, recall, F1, ontology conformance, and hallucination across subjects, relations, and objects. Baseline results with in-context learning show high ontology conformance but only moderate extraction F1 (0.25-0.35). This suggests that prompting can help models use standard relation vocabularies, but they still struggle to extract accurate triples.

5.1.5.4 Implications for the choice of generative information extraction as a system for this dissertation

Why use LLMs at all, given the caveats just summarized? The reason is their capability and flexibility concerning surface forms – exactly the challenge given by implicit, descriptive, and OOKB references. Resolving these references needs broad contextual knowledge, which only large-scale pretraining provides. For early-modern art-historical prose, there are no span-level or fine-tuned alternatives. This is not a step back from the state of the art but rather a step toward it. Generative, instruction-following extraction is currently the leading method for IE (Zhang et al. 2025e). In the DH, recognizing implicit entities is only possible through the in-context methods offered by LLMs, such as zero-shot, few-shot, and instruction-following prompting, rather than task-specific training. This is because there is not enough annotated data to train a dedicated extractor at the needed scale. Traditional methods rely on this kind of supervision, but LLMs do not. However, LLMs are still imperfect at this task. For example, Schimmenti et al. (2025) found that performance drops on the implicit formulations of sentiments, generally common in humanities texts. Because of this, the extraction process must be documented rather than simply

trusted. So, using LLMs is both necessary and imperfect, and the rest of this subsection will discuss both aspects.

The decision to go with generative IE must be made carefully, especially given observations from past studies summarized in the previous subsections: fine-tuned transformers still outperform much larger general-purpose language models on in-domain benchmarks (Zhang et al. 2025e, Gesese et al. 2026). CoT prompting brings only small or inconsistent improvements in extraction scores (Polat et al. 2025, Winkle et al. 2026). Evaluation frameworks designed for discriminative systems often misjudge generative output (Fan et al. 2024, Xu et al. 2024, Stramiglio et al. 2025), and LLMs show clear performance drops on the implicit formulations common in humanities corpora (Stramiglio et al. 2025, Tint et al. 2026). Using generative systems for IE, thus, entails additional considerations.

The main reason for this choice is that there is no real alternative. There is no pre-trained model for early modern Italian art historical prose, and no annotated corpus exists for implicit or explicit artwork entities, although these entities are a crucial component of Renaissance scholarship. Creating either requires significant manual effort at the scale required by competitive fine-tuned extractors. Building a task-specific student model from a larger teacher, as in UniversalNER (Zhou et al. 2024c), is theoretically possible but costly, and its applicability has already been attempted by Santini et al. (2022), where artwork extraction from a part of *The Lives* reported 0.55 F1 for artwork entities and 0.60 F1 for depicted subjects.

However, in comparison to the approach we aim for, the annotation scheme treats artwork references as surface spans to be linked to Wikidata, and of the 133 annotated WORK mentions in Santini et al. (2022)'s set, 108 (about 81%) resolve to `NIL` because there is no matching entry in the KB. There is no differentiation between implicit- and explicitness, which results in a heap of unresolvable mentions, which is exactly what makes the nature of historical sources: *"David who was holding Goliath"*, *"S. Margaret issuing from the belly of the Dragon"*, *"two half-length prophets"*, *"copy of the picture in which Raffaello da Urbino had formerly painted portraits of Pope Leo, Cardinal Giulio de' Medici, and Cardinal de' Rossi"*. A discriminative span-and-link pipeline cannot handle these cases because the task is not just about finding spans but about interpreting them using iconographic and art-historical knowledge that is not present in the training data or the KB. Neuro-symbolic architectures that add ontological constraints to the extraction process (Hitzler et al. 2024, Deng et al. 2021) help by enforcing schema compliance during decoding, but they assume the interpretive problem is already solved, that the mention has been found and typed, before symbolic validation happens. In a corpus where identification and typing are themselves interpretive acts, the ontology cannot be merely a post-hoc check on outputs; it must guide the generation process from the beginning. This is what prompting with

ontology-derived instructions allows, and what the Viewsari approach described earlier puts into practice.

We want to highlight that the pro and contra arguments for and against a generative approach are manifold; however, the rise in papers on generative IE shows a clear bias towards the latter (Zhang et al. 2025e). Especially their domain independence and the lack of intensive training per use case make them not only a suitable fit for our approach; another strong case comes from Dagdelen et al. (2024), who experiment with GPT-3 and Llama-2 on annotated corpora of similar size to the GT used in this dissertation. Their Llama-2 model on the doping task achieved an F1 score of 0.821, compared to 0.390 for a MatBERT-plus-proximity baseline (Walker et al. 2021) and 0.496 for a seq2rel sequence-to-relation model (Giorgi et al. 2022).

However, what remains unresolved is the task that their study also left open, particularly in the empirical sciences and the humanities. While their approach does not require the interpretive judgment required for implicit entity resolution in Viewsari, it is extremely relevant to detecting implicit artworks and resolving epistemological issues. Identifying *"Bi2Te3"* as a chemical formula is a pattern-matching task. In contrast, deciding that *"the minister of culture to the Medici court in Florence"* refers to Vasari in a specific biographical passage requires understanding historical context, authorship conventions, and the phrase's place in *The Lives*.

Fine-tuning on 400 paragraphs like these would help the model learn the basic patterns of implicit reference in the GT. However, it would not help the model handle the broader range of interpretations that come from pretraining on large collections of historical, literary, and art-critical texts. For this reason, our approach focuses more on ontological formalization and is not just about improving extraction tasks. We developed ObliquER not because it outperforms fine-tuning, but because it is the only method that can document the interpretive steps needed for implicit entity resolution in Vasari. This is possible because the model relies on broad pretraining rather than narrow domain supervision. LLMs are known to do well in this context (Zhang et al. 2025e), and this is where ObliquER is used.

The lack of transparency in pretraining data becomes a core methodological argument rather than just a background detail. When a model identifies *"the minister of culture to the Medici court in Florence"* as Vasari, it relies not only on the immediate text but also on patterns learned from sources that researchers cannot review (Li et al. 2024). This is not a typical fairness issue, but it is similar in structure: the extraction is influenced by patterns such as the over-representation of well-known figures; basically, a repetition of Vasari's biases and mistakes as if they were neutral facts. Standard evaluation cannot flag these as errors because, on the surface, they are not wrong; they are interpretively complex in ways that only provenance documentation can reveal.

When an LLM serves as an annotator, it is important to evaluate both the accuracy of IE and the contextual information provided with each annotation. This dissertation addresses these considerations by retaining uncertainty and provenance metadata within the Viewsari ontology, rather than reducing outputs to a single best label. However, provenance is not the only way to make ML systems more transparent. There is a large body of work in explainable and interpretable AI that aims to make model behavior clear during inference, including post-hoc explanations, feature attributions, and analyses of internal representations. Comparing the dissertation's approach to this related tradition helps clarify what a KG-level method can and cannot do, and where its limits are. The next section reviews explainable and interpretable AI, focusing on how it overlaps with, differs from, and complements the provenance-oriented approach discussed earlier.

In Chapter 7, we describe the extraction architecture that follows this approach, and Section 5.2.4 places ObliquER in the context of previous computational work on Vasari.[2]

5.1.5.5 A limitation that explainability cannot explain

The limitations we surveyed in Section 5.1.5.2 are not just about what generative IE spits out (sometimes with more enthusiasm than precision), or how we score those outputs. There is also the matter of accountability. While explainable and interpretable AI is the usual suspect when someone asks, 'Why did the model choose this option?' We contend that XAI and the provenance demanded by humanities scholars are answering questions from entirely different perspectives. One cannot simply change one for the other, no matter how tempting the shortcut. To clarify, we compare these perspectives along three axes, namely the target question, the granularity of explanation, and the epistemic addressee (**C3**). For ObliquER, this means that patching the provenance gap (**C2**) is not addressed by the XAI toolkit, even if it is the closest thing in the academic toolbox. Instead, we suggest a solution at the level of representation via a new modeling approach. Enter the Viewsari ontology (see Chapter 7), our answer to this conundrum. As we show in Section 10.3, the resulting design pattern – Computational Provenance (**C4**) – is an addition to the pattern catalogs.

[2] The specific prior computational work on Vasari's *The Lives*, including Santini et al. (2022) and their full evaluation setup, is discussed in detail in Section 5.2.4.

5.1.6 Explainable and interpretable artificial intelligence

The previous section concluded that XAI does not provide the accountability required for humanities extraction. This section substantiates that claim. Interpretable ML addresses feature-attribution questions regarding model internals, whereas humanities readers pose source-critical questions concerning scholarly assertions. These two approaches differ along three primary axes:

- **Target question.** XAI asks *which input features or internal computations influenced this output*; hermeneutic provenance asks *on what textual basis, under which interpretive convention, and by which agent was this claim made*.
- **Granularity of explanation.** XAI operates at the level of tokens, attention heads, or training-data exemplars; hermeneutic provenance operates at the level of source passages, prompts, model versions, and extraction runs.
- **Epistemic addressee.** XAI answers a model developer or auditor asking *why does the model behave this way*; hermeneutic provenance answers a scholarly reader asking *can I cite, contest, or extend this assertion*.

The remainder of this section walks through the principal XAI techniques in turn and shows that each of them resolves the first axis but leaves the third untouched.

In XAI, model-agnostic explanation frameworks such as Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016) and Anchors (Ribeiro et al. 2018) were developed to address this challenge. These methods build local post hoc approximations of model behavior, allowing us to examine the 'black-box' nature of individual predictions without access to the model's inner workings.

However, we want to highlight that this technical notion of interpretability differs from the epistemic interpretability discussed in this dissertation, as already introduced in Chapter 2. The technical approach answers questions like *which features influenced this model's output?*, while our focus is on questions such as *on what interpretive basis was this assertion made, by whom, and under which conditions?* Frameworks like LIME can show how individual tokens affect predictions, but they do not reveal the reasoning behind the label, the context, or the further metadata.

As generative models replace discriminative classifiers, maintaining the technical idea of interpretability becomes more difficult. This is because models like UniversalNER have open output spaces at inference time and complex, distributed knowledge that does not fit the local linear approximations used by frameworks like LIME, leading to extensions such as LIME-LLM (Mihaila et al. 2026).

Recent research has extended interpretable ML to generative models. These include CoT prompting as a form of self-explanation, in which the model articulates its reasoning before producing a final answer (Wei et al. 2023), and approaches that aim to determine training data from models (Zhang et al. 2025d, Hohensinner et al. 2026). Although these approaches represent advances in model transparency, they do not resolve the provenance problem central to humanities extraction, as they remain focused on the question of *why did this model produce this output?*, addressing the internal mechanics of a prediction, not its epistemic status as a scholarly assertion.

CoT prompting reveals the model's intermediate steps – but these steps are still generated from the same opaque dataset the model uses (Turpin et al. 2023). For instance, if the model reasons, "the Sistine Chapel is associated with Michelangelo; therefore, the master referred to is Michelangelo," we still do not know the context behind these assumptions. The scale and epistemological problems, discussed in Section 2.1, remain unsolved. Instead of just explaining the model's behavior, which can be expensive, as shown in (Hohensinner et al. 2026), the goal is to construct a formal representation of the conditions under which each statement was made, enabling tracing of outputs to their epistemic conditions. The provenance layer of Viewsari, described in Chapter 7, is designed for this purpose.

5.1.7 Generative entity linking

Recent EL methods usually use a two-step process: candidate generation and ranking, with a bi-encoder and cross-encoder architecture (Logeswaran et al. 2019, Wu et al. 2020, Ayoola et al. 2022). The bi-encoder encodes each KB entity description and matches it to the entity mention in the text using similarity measures. This step quickly finds an initial set of candidates, which is important for the next step and helps lower computational costs. In the cross-encoding phase, a cross-encoder ranks the candidates by training on the combined input text and entity description. The cross-encoder is trained to assign higher scores to matching pairs. The entity with the highest score is chosen as the final link.

Recently, entity linkers have increasingly utilized autoregressive methods based on generative models (De Cao et al. 2021, 2022, Shi et al. 2024). Some methods directly generate entity identifiers by memorizing the entire KB during training, thereby avoiding the challenge of mining negative samples, which is critical to candidate-generation performance (De Cao et al. 2021). Other approaches retain a candidate generation step but employ generative models for ranking (Zhou et al. 2024b). These methods are typically pre-trained on large corpora with a primary focus on explicit entity mentions, making them unsuitable for implicit ones.

Few studies have explored linking implicit entity mentions (Perera et al. 2016, Hosseini and Bagheri 2021), and most operate on social media text. Hosseini et al. (2018) recast implicit entity

linking in tweets as ad-hoc document retrieval, ranking candidate entities of a pre-specified type by the relevance of their user-generated reviews to the tweet through a model that interpolates term-dependency, explicit-entity, and neural-embedding features. Generally, the task is defined over an input already assumed to contain an implicit mention of a known type, from which a single entity of that type is returned, so whether a passage refers at all, and how many referents it carries, falls outside the task; Hosseini (2019) identifies precisely this skipped recognition step as a limitation that renders prior implicit linking of little practical use. Because the mention has no surface form to match, candidates are not drawn from the mention itself but from a precompiled model of the domain entities or from ad-hoc retrieval over related content, rather than from the mention string. Such models also confine these systems to entities already represented in the KB, with novel or emerging referents reported as an open failure case. Unlike these studies, our work targets longer texts in which several implicit mentions co-occur and reference must be established before it can be resolved, and it treats OOKB entities as first-class outputs, which we elaborate in Section 5.1.8.

In the DH and CH domains, EL methods, especially for historical texts, remain scarce. The Identifying Historical People, Places, and other Entities (HIPE)-2020 (Ehrmann et al. 2020) and HIPE-2022 (Ehrmann et al. 2022) shared tasks, for instance, focus on linking entities in historical news articles and consider OOKB entities, but they do not cover implicit entity mentions. In the same way, Graciotti et al. (2025) studies EL in historical music documents but looks only at explicit entity mentions. Santini et al. (2026) introduce MHEL-LLaMo, an unsupervised system for Multilingual Historical Entity Linking (MHEL). It combines a multilingual bi-encoder with an instruction-tuned LLM. The bi-encoder retrieves candidates from Wikidata, and the system uses prompt chaining with the language model for sequential NIL prediction. Candidates are chosen only if the bi-encoder's confidence is below a set threshold. This approach is based on the idea that LLMs can sometimes make mistakes in simple cases but are helpful in more complex or unclear situations. ObliquER uses a similar method for implicit EL from *The Lives*. It treats candidate generation and linking as separate steps, with the LLM handling inferences, expanding the capabilities of a pure knowledge-base lookup. ObliquER grounds its process in the provenance-aware Viewsari schema and uses dedicated OOKB typing. The two-step paradigm is necessary but not sufficient for the Vasari case, since there is no surface-form anchor for implicit mentions in candidate generation. Correspondingly, when ranking assumes the correct referent is in the candidate pool, this is not possible for OOKB entities by definition. The Viewsari schema addresses both gaps by representing these entities as first-class graph nodes rather than NIL labels, a design requirement that falls outside the scope of any retrieve-then-rank architecture and is discussed in Section 5.2.5.

5.1.8 Of the long tail and resolving out-of-knowledge-base entities

LLMs struggle to handle long-tail knowledge. Kandpal et al. (2023) shows that model accuracy on fact-based questions is linked to how often relevant information appears during pretraining, suggesting an imbalance between parametric and distributional knowledge. In low-resource situations, models might produce information that sounds correct but is hallucinated. The long-tail/OOKB distinction established in Chapter 4 is directly relevant to the extraction setting discussed in this dissertation, since Vasari's *The Lives* includes many entities that are long-tail by any standard, such as unknown or lost artworks. Many of the entities that ObliquER aims to detect are not only rare but also absent from resources such as Wikidata, and, to make matters worse, many of these figures are mentioned implicitly, making the setting even more difficult for extraction systems that rely on explicit mentions. The modeling and evaluation strategies used to address these challenges are discussed in Chapter 7.

The necessity of explicit knowledge representation within symbolic structures has been challenged by research that treats LLMs as implicit KBs. Petroni et al. (2019) demonstrate that pretrained transformer models encode factual knowledge in their parameters, which is recoverable via cloze-style probing. This research suggests that parametric memory functions as a sort of KB, thereby blurring the boundary between in-knowledge-base and OOKB entities. However, this representation is neither guaranteed to be complete nor epistemically transparent, and performance declines sharply for certain relational patterns and rare entities (Kandpal et al. 2023). Consequently, the encoding of parametric knowledge in LLMs, even if partially retrievable, does not eliminate the OOKB problem but shifts it to statistical sparsity.

The work by Graciotti et al. (2025) not only addresses KE for CH applications in historical sources but also identifies this long-tail problem as unavoidable – in these contexts, OOKB entities constitute the norm rather than anomalies. However, those rare and domain-specific entities, such as historical figures and events, significantly degrade model performance.

This shows that even with recent progress, LLMs still struggle with OOKB entities and have broadened the range of long-tail knowledge. In autoregressive methods, some solutions include setting a threshold (Ayoola et al. 2022) to decide when no candidate fits, or letting the generative model choose when no match is found. But these approaches assume each mention fits exactly one category (NIL or linkable) and miss cases where several mentions in a corpus refer to the same entity that is not in the KB. Ideally, an EL approach should not only identify these cases but also cluster them. Möller and Usbeck (2024) address this detection-and-clustering problem using only the KB. They cluster OOKB mentions by comparing pairwise similarities that use both mention embeddings and the knowledge-base embeddings of nearby linked entities. Their results

show that context from the KG helps more in distinguishing known entities than in resolving coreference among unknown entities. This research matters here because, like ObliquER, it assumes that entity extraction cannot depend on a fixed set of canonical referents. This is especially important for historical sources, where, as Graciotti et al. (2025), Santini et al. (2026) point out, ambiguity, incomplete records, different naming practices, and disagreements among historians are normal features rather than mere noise. So, resolving OOKB entities has become an interpretive task. These entities are not just errors to fix; they can show the limits of both symbolic and parametric knowledge systems. This is the configuration ObliquER operates in, since we build this pipeline for a corpus in which those entities are structurally predominant, making their identification interpretive rather than referential. The absence of a canonical identifier must be represented explicitly rather than treated as an extraction failure, resulting in a set of representational requirements which we elaborate on in Section 5.2.5.

5.1.9 Best of all worlds? Project example: Computer-assisted semantic text modeling

Historical texts often hide meaning within layers of storytelling. When extraction systems are configured to produce data triples, they risk turning these layers into context-free statements. The CASTEMO (Computer-Assisted Semantic Text Modelling) framework takes a new approach to the challenges of extracting historical knowledge. Instead of pulling out isolated 'facts' from texts, CASTEMO suggests modeling the source itself first to overcome the challenge of reducing complex narratives to simple, disconnected data points (Shaw et al. 2026).

They point out an overemphasis on the collection of facts, whereas humanities scholarship thrives on details and context, which, when taken out of their original place and reduced to simple statements, lose their dimensions completely. While this makes quantitative analyses easier, it puts interpretation and 'fact' on the same level, and important details about the process and background of where the data came from are absent from the final form.

For example, consider this passage about Michelangelo:

> Michelagnolo took that drawing and with a thicker pen outlined one of those women with new lineaments [...] That sheet is now in my possession [...] In the year 1550 [...] he showed it to Michelagnolo, who recognized it and was pleased to see it again, saying modestly that he knew more of the art when he was a boy than he did at that time, when he was an old man [...].

A conventional KE pipeline might produce statements such as:

- Michelangelo corrected a drawing by a fellow pupil.
- The drawing was copied from works by Ghirlandaio.
- Vasari possessed the sheet.
- Michelangelo acknowledged the drawing in 1550.

Although these statements are accurate, they remove the layers of the source, as it does more than just list events; it presents them within a mix of admiration, reflection, and self-positioning. CASTEMO solves this by modeling manual annotations in the texts based on how statements relate to one another and by separating different levels of meaning, namely what is directly stated, what is clarified by editors, and what is inferred by scholars. It also notes whether something is an assertion, a question, a denial, or a condition, and how certain it is.

The role of LLMs in such a framework is not self-evident. When applied only to extract information, the results will masquerade as simple facts. However, when used as annotators within a system that tracks the source of information, this role changes. Instead of just finding facts, the LLMs encode structured claims within the text. This is not just a technical change but a change in how we understand knowledge generation. Entities are not just identified but rebuilt from the story, so the results keep the different layers of meaning and recognition. This dissertation sees CASTEMO as an important step away from the fact-focused approach of many current systems. It shows that detailed modeling and attention to nuance can go hand in hand.

Viewsari, like CASTEMO, models the source before extracting facts, but it takes a different approach in two important ways for provenance. While CASTEMO uses manual annotations to store statements in a custom JSON format, Viewsari is an OWL-based ontology built on established Semantic Web standards, such as PROV-O for provenance (Lebo et al. 2013) and the International Federation of Library Associations and Institutions (IFLA) family (Riva et al. 2017) for detailed bibliographic structure. Additionally, CASTEMO aims to capture epistemic level and certainty as editorial judgments on statements coded by humans; ObliquER, however, highlights interpretation by an LLM and treats its CoT reasoning as additional provenance objects. In this way, Viewsari builds on CASTEMO's idea that the conditions under which information is produced should be part of the data itself, not just a footnote. It adapts this idea to a context where the interpretive agent is generative, so provenance must cover both the source's context and the reasoning behind each extraction.

5.2 Knowledge engineering

Knowledge engineering is the formal process of creating frameworks to represent knowledge. It is the backbone of KB systems, generally described as a modeling activity that produces computer systems capable of recreating human decision-making comparable to that of a domain expert (Studer et al. 1998). In the following, we focus, in particular, on this for creating ontologies, so ontology engineering, as the baseline for KG construction on a given schema.

5.2.1 Advancements in ontology engineering methodologies

Ontology engineering extends formal ontology modeling to encompass iterative, requirements-driven processes, with key aspects including modularity, engagement with domain experts, and validation through real-world use cases. In research domains such as the DH, these methodologies offer structured approaches for integrating formalizations with domain-specific expertise. Multiple methodologies have been developed to facilitate ontology development. Contemporary ontology engineering has transitioned from monolithic, waterfall-style modeling to iterative, reuse-oriented, and collaborative approaches. Among these, XD is a particularly influential framework (Presutti et al. 2009).

Viewsari is continuously developed using the XD methodology, in which design choices are informed by CQs derived from user personas representing distinct use cases. The rationale for preferring XD over alternative methodologies is detailed in Section 7.3.

XD incorporates principles from agile software engineering and emphasizes the systematic reuse of ODPs. ODPs provide modeling solutions for recurring design problems in ontologies and serve as foundational building blocks for ontology modules (Gangemi and Presutti 2009). XD supports iterative development cycles driven by CQs and test cases. Modeling decisions are validated incrementally, and small, reusable pattern-based modules are used instead of large, monolithic schema construction.

As a user-centered design method, XD emphasizes the active involvement of domain experts in guiding ontology design decisions, either directly or indirectly (Presutti et al. 2009). Domain experts participate throughout the process, while ontology engineers, following software engineering practices, capture requirements as user stories that illustrate real-world scenarios for potential ontology content. Personas may complement user stories by representing extreme stereotypes of potential users (Junior and Filgueiras 2005). User stories form the basis for CQs, and together they provide the foundation for identifying and creating ODPs (Presutti et al. 2009).

At different abstraction levels, ODPs construct six groups, as described in detail by Gangemi and Presutti (2009). Selected ODPs have the same functionality as software templates, offering adaptability for specific domains and application scenarios (Blomqvist et al. 2016). ODPs are accessible through repositories such as `OntologyDesignPatterns.org`. For this dissertation, only content patterns are considered relevant. These patterns encode conceptual design solutions specific to domain classes and properties within ontologies (Gangemi and Presutti 2009). Employing modular patterns helps maintain sectioned solutions, offering smaller frameworks for particular use cases that are easier to manage (Blomqvist et al. 2016). This approach relates to a similar endeavor by Shimizu et al. (2019), who introduce the ongoing development of MODL, a library for modular ontology design that aims to abstract and harmonize patterns across different ontologies.

In the field of CH, such practices have been implemented, for example, in the Knowledge Graph of the Italian Cultural Heritage (ArCo) (Carriero et al. 2019); its full pattern-based methodology and unit-testing tool, which pair each CQ with a SPARQL-executable answerability check against the populated graph, are documented in Carriero et al. (2021) and provide the methodological precedent for the metric catalog of Section 8.3. Nevertheless, widespread adoption remains limited, and no comprehensive standard exists for integrating diverse ontologies within DH projects (Meroño-Peñuela et al. 2014, Doerr 2003, Wang and Zhang 2025). In art history, patterns have been used in the ICON ontology to describe the interpretive processes of artworks (Sartini et al. 2023) or to formalize digital collections of artworks (Ahola et al. 2025). More on these ontologies is given in the following Section 5.2.2.

5.2.2 Ontologies in the digital humanities

The representation of CH and historical artifacts in formal systems has been a central concern in the DH for more than two decades (Doerr 2003, Meroño-Peñuela et al. 2014, Hyvönen 2026). Various ontologies and catalogs of best practices have emerged from these efforts. This section presents a summary of ontologies used in Viewsari, including the IFLA and Semantic Publishing and Referencing (SPAR) families for bibliographic content in Section 5.2.2.3, the Provenance Ontology (PROV-O) for provenance chains in Section 5.2.2.2, and several domain-specific extensions related to the Viewsari ontology, as well as a comparison to the CIDOC CRM in Section 5.2.2.1 and the Historical Context Ontology (HiCO) in Section 5.2.2.4 for why they were not used. These ontologies show an overview of how to collectively enable the structured modeling of heterogeneous DH sources.

These standards are the formal foundation for Viewsari. Rather than building a custom schema, this dissertation's ontology brings together and extends PROV-O, the Web Annotation Ontology,

FaBiO/FRBR, and the Document Ontology. This approach supports interpretive claims from Vasari's work and connects them to both their textual sources and the processes that produced them. Key modeling choices, like treating co-occurrences as main entities, using a three-layer bibliographic structure, and adding shortcut properties for faster queries, are explained in Chapter 7. This section reviews existing ontologies and the modeling traditions that shape Viewsari. We also contrast other existing ontologies, such as CIDOC CRM, with Viewsari, demonstrating key differences and gaps where they are insufficient to accommodate the requirements for representing interpretive processes in generative IE from historical sources.

5.2.2.1 CIDOC CRM as a comparison to Viewsari

CIDOC CRM provides an ontological infrastructure for integrating and exchanging heterogeneous CH information. Developed by the CIDOC Documentation Standards Working Group of ICOM, it formalizes event-centric scenarios and establishes a semantic framework for interoperability among museums, archives, libraries, and other institutions. This fundamentally event-centric model extends the view of CH artifacts from static objects to interconnected items, temporal entities, and actors' participation in events (Doerr 2003).

Over the years, it has become the de facto standard for modeling CH knowledge, the most used ontology in the CH domain, and an ISO standard (ISO 21127) since 2006 (Ranjgar et al. 2022), so not using it for projects at the intersection of the DH and the semantic web often requires rigorous justification. While the event-centric structure supports the explicit representation of production events, detailed object biographies, roles of actors, temporal spans, and material transformations, its core model does not account for scholarly handling of the data; for this, several extensions exist, including CRMinf, a global schema for integrating metadata about interpretation, argumentation, and inference-making in descriptive and empirical sciences. Nevertheless, there is no module for interpreting or providing provenance of extracted information (Doerr et al. 2011, 2023).

Tzitzikas et al. (2022) empirically confirms this gap in their survey of ML approaches applied to CIDOC CRM, showing that IE from text lacks the epistemic provenance of the extraction process itself. This means that the extracted outputs are treated as data entries rather than as documented interpretive acts.

This design reflects a key ontological commitment, establishing that CH knowledge is best understood as the result of historical processes. Consequently, objects are not described in isolation. Instead, they are placed within networks of events such as production, use, transfer, and documentation. For this, CIDOC CRM offers several core classes, including:

- `E22 Human-Made Object`

- `E39 Actor`
- `E5 Event`
- `E12 Production`

In practice, the application of CIDOC CRM to CH artifacts is manifold; let us look at the production of a painting in the Italian Renaissance:

```
Leonardo da Vinci painted the Mona Lisa in Florence around 1503.
```

In CIDOC CRM, this production of a painting by Leonardo da Vinci becomes an event structure, providing the contextual information based on a `crm:E12_Production` event and related metadata on, for example, `crm:P108_has_produced` with the result `:Mona_Lisa` and `:Mona_Lisa a crm:E22_Human-Made_Object` (Doerr 2003).

Although the ontology allows for incomplete knowledge, it primarily records factual statements about an object's history, with the object being an actual referent in the real world. It shows what is claimed to have happened, but not who made the claim or under what conditions. This limitation is well known in the field of CH, especially when the main interest is scholarly discussion of an object, such as attributions, debates over dates, iconographic interpretations, and other claims (Doerr et al. 2011). As mentioned earlier, extensions such as CRMinf have been added to the core ontology to address this. CRMinf models propositional attitudes of human agents, including belief, inference, and observation. It was designed for curatorial interpretation in CH documentation, where a human expert attributes a work, dates an artifact, or proposes a reading (Doerr et al. 2023).

A further extension, CRMdig (Doerr et al. 2016), provides classes for digital machine events and formal derivations, making it the closest existing vocabulary for documenting computational extraction processes. So, these developments are closely related to the topics in this dissertation, but they solve different problems. In particular, CRMinf is designed to model scholarly disagreements over factual claims about cultural objects – that is, real-world referents. In contrast, our work focuses on modeling the provenance of interpretive claims made by a generative information-extraction system based on textual mentions in historical sources. The difference has already been discussed in Chapter 2 and will be discussed in detail in Section 9.2.

Viewsari uses PROV-O because, compared to CRMinf's focus on human-made attributions, it has been specifically developed to track provenance in computing processes, which better matches what a generative extraction pipeline produces when extracting textual mentions from historical sources. CRMinf is based on `I2 Belief`, `I4 Proposition Set`, and `I5 Inference Making`, which assume a belief-holder who accepts a proposition on the basis of reasons. This works for

human argumentation but does not fit LLM generations, since an LLM has "[...] no notion of truth or falsehood because it lacks the means to exercise these concepts in anything like the way we do" (Shanahan 2024). The details of this modeling approach are discussed in Chapter 7. Viewsari, thus, addresses a structurally different problem, where evidence is not an observation but a text span, a specific sentence in a specific paragraph from which an entity was extracted.

However, these ontologies are not mutually exclusive and, in application scenarios, can work together to harmonize different perspectives on cultural artifacts. CRMinf and Viewsari are not competitors; they address different epistemic agents, different evidence structures, and different knowledge gaps. In application scenarios, they can work together to harmonize different perspectives on cultural artifacts. PROV-O could record how claims are generated by an extraction pipeline, while CRMinf tracks how scholars later discuss, challenge, or support those claims. For example, if an art historian disagrees with an attribution made by our extraction pipeline, this would be represented as an `I2 Belief` over an `I4 Proposition Set`, where the propositions are PROV-O-annotated entities produced by the pipeline. This dissertation focuses on the first layer, but its approach allows for later integration with the second.

5.2.2.2 PROV-O for provenance activities

PROV-O is the W3C standard for representing provenance information on the web. It formalizes the PROV data model (PROV-DM) in OWL, enabling machine-readable descriptions of the relationships among entities, activities, and agents involved in creating data or knowledge. We use PROV-O as the foundation for the Viewsari ontology because it focuses on three fundamental classes: `prov:Entity`, which encompasses physical, digital, or conceptual things; `prov:Activity`, denoting processes that act on entities; and `prov:Agent`, referring to persons, organizations, or software responsible for activities. These classes connect through core properties such as `prov:wasGeneratedBy` and `prov:wasDerivedFrom`.

PROV-O is designed to be domain-neutral, so it can be used across many domains, including scientific workflows, digital curation, and knowledge production systems. For example, imagine an extraction process that finds entities in a document:

> *Giorgio Vasari, the minister of culture to the Medici court in Florence, blessed the Pope with a gold-crested angel in the chapel ceiling.*

leading to

```
[Giorgio Vasari]<PERSON>, [gold-crested angel]<ARTWORK>,...
```

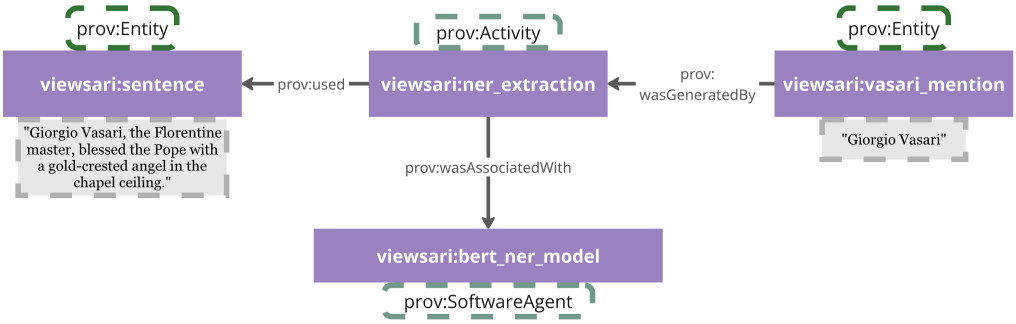


Figure 5.1: PROV-O representation of a named entity extraction step applied to a sentence from Vasari's *The Lives*. Solid purple boxes denote instances in the `viewsari` namespace; dashed green outlines denote the PROV-O classes to which those instances belong.

In PROV-O, this could be formalized as shown in Figure 5.1. Viewsari uses this when applying the pattern to generative extraction, where the agent is a language model and, in more detailed modeling, the activity includes a CoT trace as additional provenance. PROV-O enables dedicated documentation of workflows, assignment of responsibility, and reproducibility of computational processes. However, it does not specify domain semantics or interpretation types, nor does it distinguish between different epistemic criteria. It explains how something was created, but not how it should be evaluated epistemically. This is where Viewsari will step in, as discussed in Chapter 7.

5.2.2.3 Representing bibliographic content with the IFLA and SPAR ontologies

IFLA published FRBR in 1998, introducing a conceptual model for bibliographic information. The best-known part is the Group 1 hierarchy, which separates Work (the intellectual or artistic content), Expression (how that content is conveyed), Manifestation (the set of carriers with the same content and production plan), and Item (a single copy of a manifestation) (Tillett 2005, IFLA Study Group on the Functional Requirements for Bibliographic Records 2009). Later, this four-level structure was merged with Functional Requirements for Authority Data (FRAD) and Functional Requirements for Subject Authority Data (FRSAD) to create the IFLA Library Reference Model (LRM) (Riva et al. 2017, IFLA Study Group on the Functional Requirements for Bibliographic Records 2009). LRM keeps the WEMI structure but, among other aspects, adds a top-level superclass, treats Place and Time-span as first-class entities, and defines Agent as a superclass for Person and Collective Agent. It also focuses on five user tasks for library domains: find, identify, select, obtain, and explore. LRM is backward-compatible, so users can omit entities they do not need and still remain compliant with the model.

Viewsari intentionally uses FRBR instead of LRM. The Bibliographic Ontology (FaBiO), which is based on FRBR and part of the SPAR Ontologies suite (Peroni and Shotton 2018), provides a formalization of WEMI and supports different publication types. The main reason for this choice is that the application scenario for the Viewsari ontology lies outside the library domain, and it is unnecessary to support diverse library-specific user tasks. Authority data, so far, does not play a central role in Viewsari, and the WEMI hierarchy is the core component for representing the editions of Vasari's work and contextualizing provenance information for the extracted content. This makes it possible to represent *The Lives* as a work with several historical editions and translations, down to specific passages in printed versions where interpretive claims are made. The extra features LRM adds beyond WEMI (Res, Nomen, first-class Place and Time-span, and consolidated authority data) are useful for full library cataloging but are not needed for extraction and provenance in this project. Since the main WEMI hierarchy remains in LRM as `LRM-E2` through `LRM-E5`, switching to an LRM-based ontology in the future would only require changing namespace prefixes, not a full redesign. This is planned as future work (Section 12.3). Hence, we opted for FaBiO because it is well-maintained and used in other DH KG projects (Carriero et al. 2020, de Berardinis et al. 2023). More on this follows in Chapter 7.

5.2.2.4 Diverging interpretive claims in humanities texts with HiCO

According to Daquino and Tomasi (2015), HiCO is a conceptual ontology that represents contextual information about CH objects. HiCO formalizes the interpretive processes used to produce historical knowledge, especially when working with full-text sources like TEI-encoded scholarly editions. The ontology notes that information in CH artifacts is rarely based on neutral facts and that it is necessary to capture the outcomes of interpretive activities, such as interpretation, transcription, inference, and scholarly argumentation. HiCO meets the need to represent content from full texts by formalizing contextual details such as events, people, and relationships, while also modeling provenance and authoritativeness. To achieve this, HiCO builds on and extends PROV-O by incorporating provenance modeling into historical interpretation. The main class in HiCO is `hico:InterpretationAct`.

The `hico:InterpretationAct` describes how an agent extracts, conceptualizes, and formalizes information from a source. This approach helps avoid logical inconsistencies from conflicting statements. Each interpretation is linked to a specific scholarly work and can be traced back to a responsible agent. One innovative feature of HiCO is its modeling of authoritativeness. With SWRL rules, interpretations can be marked as "authoritatively based." HiCO does not define truth; instead, it supports rule-based interpretation within a community's methods. This shifts the focus of KGs in historical research toward documenting how the past is interpreted. Figure 5.2 illustrates this structure on an example sentence from Vasari's *The Lives*, with a human annotator

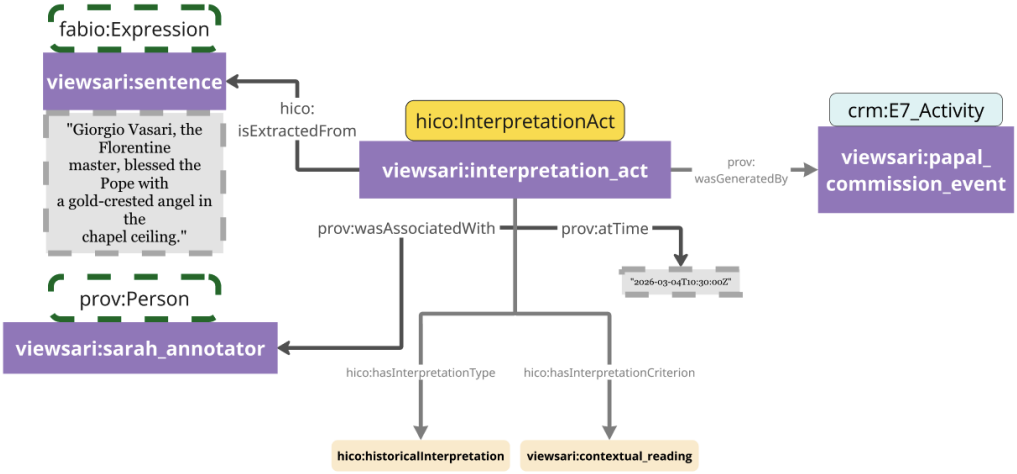


Figure 5.2: The `hico:InterpretationAct` is the single pivot of the representation: it is extracted from the source sentence (a `fabio:Expression`) via `hico:isExtractedFrom`, associated with the responsible human agent via `prov:wasAssociatedWith`, characterized by an interpretation type and criterion, and timestamped via `prov:atTime`. The claim produced by the act, here a `crm:E7_Activity` representing the papal commission event, is linked to the act via `prov:wasGeneratedBy`. Dashed outlines denote classes color-coded by namespace (green for PROV-O and FaBiO, amber for HiCO, blue for CIDOC CRM); solid purple boxes denote instances in the `viewsari:` namespace, and amber literal-style boxes hold the interpretation type and criterion values.

interpreting the passage as an instance of a papal commission event. This fully matches our hypotheses in Chapter 2, with the limitation that it is restricted to representations of human-made processes, because the interpretations made by language models are not as transparent as HiCO requires.

Thus, HiCO acts as a conceptual model for Viewsari, not as a vocabulary to reuse directly. A generative language model differs from a human annotator in lacking stable beliefs, institutional accountability, and transparent reasoning (Shanahan 2024), as HiCO's provenance module requires (Bruno et al. 2025). Using `hico:InterpretationAct` for generative output would misrepresent the model's intention. Furthermore, HiCO focuses on representing scholarly edition work, such as transcription and attribution, rather than on the provenance of generative extraction. The detailed interpretation types in HiCO (such as subclasses for editorial inference) do not align well with the operations in ObliquER, where interpretation is a prompted extraction rather than an editorial change in the tradition of scholarly editions. In Viewsari, we adapt HiCO's main idea that interpretive acts, not just their results, should be modeled in humanities KGs, but remodel its provenance layer directly on PROV-O rather than HiCO's version. The two ontologies are sufficiently close that Viewsari could, in the future, add HiCO-style interpretation classes for human scholarly annotations on top of the generative extraction layer (Bruno et al. 2025).

5.2.3 Examples of knowledge graphs in the digital humanities

This section reviews several KGs from the DH that are relevant to this dissertation. Examples include the Vespasiano da Bisticci correspondence KG in Section 5.2.3.1, Polifonia for musician networks in Section 5.2.3.2, Odeuropa with the European Olfactory Knowledge Graph (EOKG) in Section 5.2.3.3, and Adaptive Text-to-RDF for Cultural Heritage (ATR4CH), a systematic five-step methodology for LLM-based KE in Section 5.2.3.4. Although these projects offer valuable methodological foundations, the Viewsari project aims to extend and adapt these methods to better handle complex interpretive layers and to model provenance in historical texts.

Further relevant KGs include ArCo (Carriero et al. 2019), the Sampo network (Hyvönen 2026, Ikkala et al. 2022), and IICONOGRAPH (Sartini 2024). Although highly relevant for contextualizing the present work, a systematic examination of their modeling principles and architectural design would extend beyond the scope of this dissertation and is therefore addressed in the broader literature. Quantitative comparison *between* digital-humanities KGs at the structural level – triples, instance counts per class, and ontology-coverage indicators for KG quality (see, for example, Hassan et al. (2025)) is similarly out of scope, but the same structural-metrics axis grounds the population-side evaluation of the Viewsari KG in Section 8.3.

5.2.3.1 A knowledge graph of Vespasiano da Bisticci's letter correspondence

The Vespasiano da Bisticci Letters project is a case study based on the HiCO that we introduced in Section 5.2.2.4. The goal is to create a provenance-aware KG from the TEI-encoded scholarly editions of Bisticci's letters. The KG is formally grounded in FRBR, HiCO, and PROV-O (Tomasi 2012, 2013). Instead of converting TEI markup directly into RDF triples, the project adds an interpretive layer between statements. Each of these statements, which is taken from the text (for example, "Person X sent Letter Y on Date Z"), is connected to a specific `hico:InterpretationAct`. It is important to note that, as we have highlighted for use cases in HiCO, interpretations were not created automatically; they were already included in the TEI by human annotators.

Figure 5.3 shows the interrelation of letters (`fabio:Letter` and time-based aspects, connected to interpretations based on roles and agents. For instance, this would enable the formalization of annotations in which sending a letter appears as a role a person holds at a specific time in relation to a document. This event-based method aligns with historical research practices and helps prevent oversimplification of the past. Because of this, the KG presents not only the letters but also how modern scholars interact with these correspondences. This project is a relevant precedent for

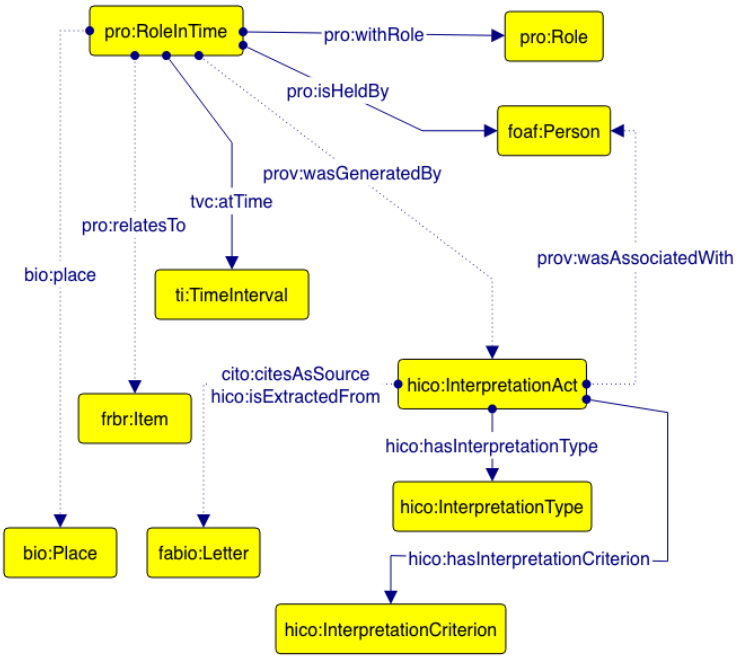


Figure 5.3: Example of how HiCO integrates with bibliographic and biographical vocabularies in a letter-corpus KG. A `pro:RoleInTime` binds a `foaf:Person` to a `pro:Role` over a specific `tvc:TimeInterval` and `bio:Place`, and is itself the product of a `hico:InterpretationAct` that records how a scholar derived the role assignment from a specific `fabio:Letter`. The interpretation act carries its own type and criterion (`hico:InterpretationType`, `hico:InterpretationCriterion`) and cites the source letter via `cito:citesAsSource` and `hico:isExtractedFrom`. Taken from Tomasi (2013).

Viewsari because both treat the source corpus as something to be interpreted rather than merely as a collection of facts. However, Viewsari uses a generative IE system as an interpretive agent, whereas the da Bisticci project relies on existing data from human annotators. We discussed this distinction in Chapter 2, which also explains why the ontology in Viewsari differs from the approaches used here.

5.2.3.2 The Polifonia Knowledge Graph and the MEETUPs ontology

The Polifonia Knowledge Graph was created as part of a major European research project focused on representing European musical heritage, and one key result is the Polifonia Ontology Network (PON), which addresses fragmentation and poor interoperability between existing music heritage collections. The ontology is the backbone of the KG, bringing together a wide range of resources related to musical heritage (de Berardinis et al. 2023, Renda et al. 2023).

PON is a modular ontology, in which the modules have different core concepts, such as music metadata, music representation and notation, provenance and source information, and more. These modules were created using the XD ontology design method. The team collected requirements as CQs, informed by user stories and pilot projects, ensuring that the KG aligns with real research and exploration tasks undertaken by musicologists and CH experts (Presutti et al. 2009, de Berardinis et al. 2023).

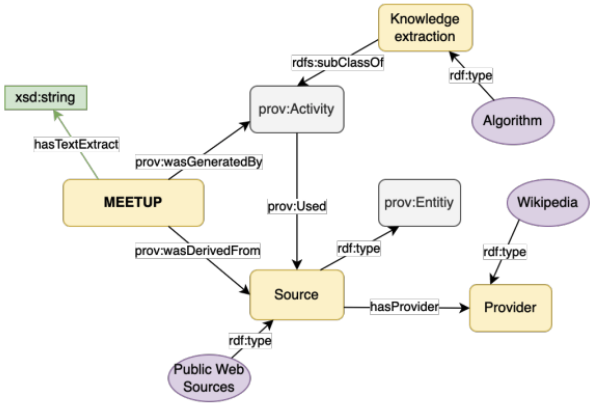


Figure 5.4: The MEETUPs event model grounded in PROV-O. A `MEETUP` is modeled as a subclass of `prov:Activity`, produced by a `Knowledge extraction` activity (typed as an `Algorithm`) and derived from a `Source` (a `prov:Entity`) whose provider is classified as Wikipedia or another public web source. The extracted text content of the meetup is attached as an `xsd:string` literal via `hasTextExtract`. Adapted from Tirado et al. (2023).

As another part of the Polifonia ecosystem, the Musical Meetups module (MEETUPs), which serves as the baseline for network analysis, displays historical meetings and interactions between musicians and other cultural figures. The MEETUPs ontology defines a historical meetup to describe events like collaborations, performances, educational meetings, or chance encounters in the European music scene. Each meetup is described by who was there, where and when it happened, and the purpose or type of event, such as a business meeting, public celebration, music-making, or education. Figure 5.4 shows how a `MEETUP` instance is represented in PROV-O terms: it is a subclass of `prov:Activity`, generated by a knowledge-extraction process (typed as an algorithm) and derived from a `prov:Entity` source, which in turn carries a provenance type indicating whether it came from Wikipedia or another public web source. The team used NLP, EL, and event classification to identify people, places, times, and meetup types in Wikipedia texts, and formalized them in the KG based on PON and MEETUPs. This method enables

large-scale analysis of historical social networks, allowing researchers to identify collaborations and patterns of cultural exchange among musicians (Tirado et al. 2023). The MEETUPs module is an important methodological precedent for Viewsari, as its event-focused approach to modeling meetups as structured encounters between cultural figures parallels the way Viewsari tracks co-occurrences among people in Vasari's *The Lives*. MEETUPs uses automatic extraction from unstructured text instead of relying on human TEI annotation, as the Vespasiano project does. This approach is similar to the generative extraction pipeline at the heart of ObliquER. The Polifonia KG is revisited in Chapter 10 as a case study for generalizing the Viewsari ontology approach.

5.2.3.3 The European Olfactory Knowledge Graph for the representation of immaterial cultural heritage

The EOKG is the main result of ODEUROPA, a European research project focused on the history and culture of smells, merging information about them, related events, objects, and emotions, along with their historical backgrounds, using data from texts and images. These references are identified and sorted using ML, then organized with the Odeuropa ontology. Each smell event or reference is described using RDF triples that connect the smell to its descriptors, related practices, and details such as time, place, and emotions (Lisena et al. 2021, 2022).

The formal core is the ODEUROPA Ontology, built on CIDOC CRM to connect smell-related events and entities, covering both physical sources of scent and related descriptions, experiences, and interpretations of smells. Those are references found through automated analyses of texts and image sources, and the ontology formalizes descriptors like 'woody' or 'animalistic,' as well as information about time, place, emotions, and the people involved, in such descriptions and smell experiences. Controlled vocabularies enable the expression of these concepts across languages, enabling cross-lingual searches and comparisons of sensory heritage.

Figure 5.5 shows how this modeling strategy works for a single smell event. A `Smell Interaction` (`od:L15`) happens in a gallery and is driven by an `Olfactory Experience` (`od:L13`), where the Author notices a specific `Smell` (`od:L11`) described with several attributes: *dusty*, *sweeter*, *musty*, and *old*, using `crm:E13_Attribute_Assignment`. The smell comes from a `Smell Emission` (`od:L12`) that can be traced to one or more sources, such as woodwork, leather, tapestry, camphor bags, or a potpourri production event, and is located in real historical places like Knole or another old house.

EOKG demonstrates that KGs can capture immaterial cultural aspects, such as smell, which are often difficult to represent in structured data. The combination of semantic technologies with ML

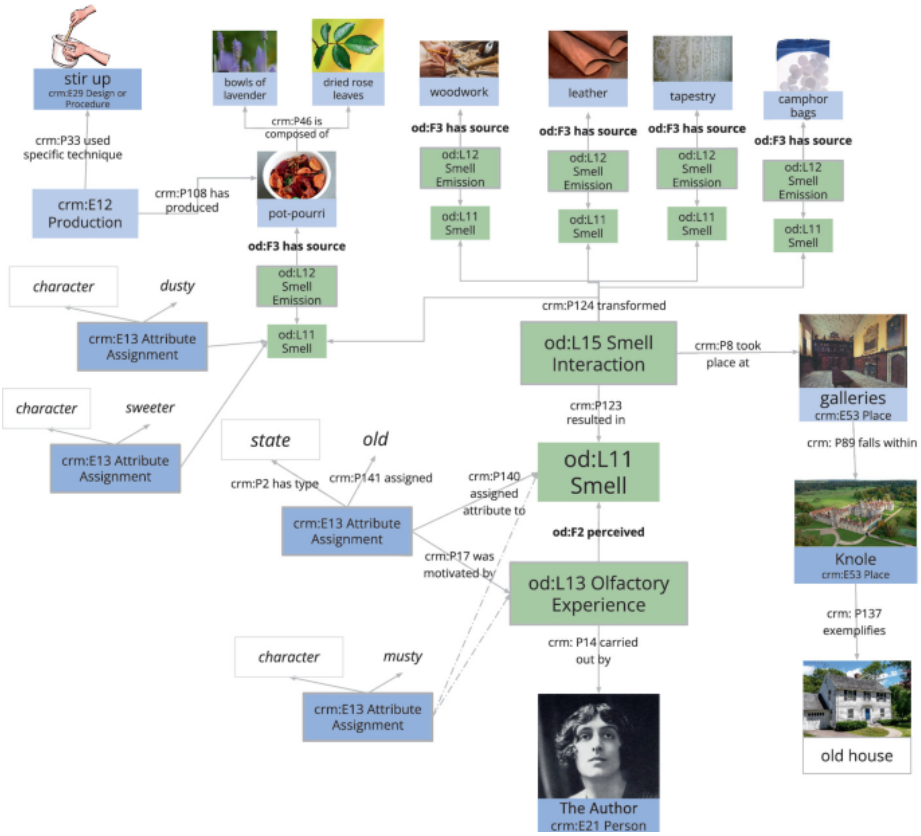


Figure 5.5: A smell event modeled in the ODEUROPA ontology. A `Smell Interaction` (`od:L15`) is the result of a `Smell Emission` (`od:L12`) from one or more material sources (woodwork, leather, tapestry, camphor bags, potpourri), takes place in a gallery of an old house (here Knole), and motivates an `Olfactory Experience` (`od:L13`) in which the Author perceives a `Smell` (`od:L11`) characterized by multiple `crm:E13_Attribute_Assignments` (*dusty*, *sweeter*, *musty*, *old*). The model extends CIDOC CRM with the `od:` namespace for olfactory-specific classes and properties. Adapted from Lisena et al. (2022).

enables researchers to study how people in the past described, experienced, and valued smells in greater detail and in a semi-automated manner.

This is relevant to our approach because, like MEETUPs, the EOKG uses automated extraction from a variety of textual and visual sources rather than relying on human annotation. Moreover, it treats interpretive descriptors such as *dusty* and *animalistic* as important attribute assignments, similar to how Viewsari treats co-occurrences as entities with their own provenance. We elaborate on this structural similarity in Chapter 10, where EOKG is used as a case study to demonstrate how the Viewsari computational provenance pattern can be applied beyond Vasari's *The Lives* to other DH KGs, thereby unifying approaches and improving interoperability.

5.2.3.4 Authenticity assessments of interpretive claims based on HiCO with ATR4CH

Schimmenti et al. (2026) introduce ATR4CH as a five-step method that uses LLMs to extract knowledge from CH texts. This approach was tested with a case study on authenticity debates. The resulting KG is based on the Scholarly Evidence-Based Interpretation (SEBI) ontology, which builds on the HiCo ontology to represent evidence for interpretations (Pasqual et al. 2024).

SEBI is organized around four main patterns, namely document metadata, scholarly opinions, evidential features (details that support assessments, including criteria and their positive or negative value), and alternative hypotheses (different claims about who created the item, when, where, or why). Each interpretation is shown as a `hico:InterpretationAct`, linked to its author with `prov:wasAttributedTo` and to supporting evidence with `sebi:support`. RDF-star allows multiple, even conflicting, claims to be included without causing logical issues.

SEBI and Viewsari differ in their focus. Where SEBI assesses the authenticity of claims, Viewsari tracks the provenance and traceability of claims made by LLMs. SEBI separates claims into three subcategories and includes a document authenticity layer that classifies documents as `sebi:Forgery`, `sebi:Authenticity`, `sebi:FormalForgery`, or `sebi:ContentForgery`. To represent conflicting claims, these are wrapped in RDF-star-quoted triples that can be linked to a specific scholarly act. The interpretation-act layer wraps these quoted triples in a `hico:InterpretationAct`, attributed to a `prov:Agent` and supported by one or more evidence nodes. Figure 5.6 shows this structure in action for Lorenzo Valla's fifteenth-century interpretation of the Donation of Constantine, where Valla's negative evaluation of the document's philological features supports classification as a `sebi:Forgery`.

The underlying extraction pipeline based on LLMs has six steps: metadata extraction, opinion attribution (using GliNER for NER (Zaratiana et al. 2024)), EL to Wikidata, opinion classification,

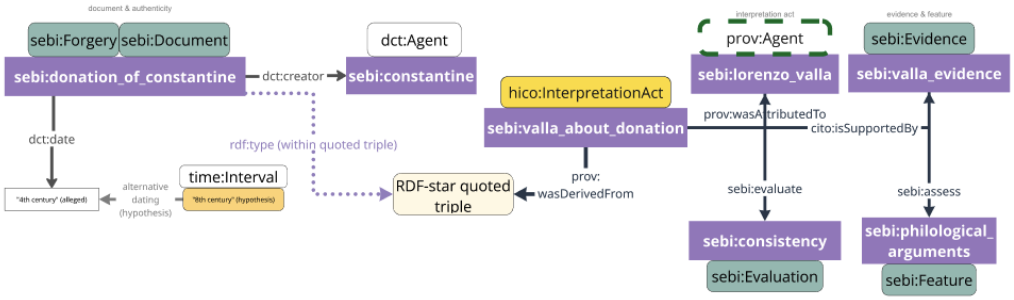


Figure 5.6: Harmonized view of the SEBI ontology instantiated on Lorenzo Valla's interpretation of the Donation of Constantine, combining SEBI's three subgraphs into a single representation. The evidence decomposes Valla's argument into its two orthogonal components, including a `sebi:Feature` being examined (philological arguments) and a `sebi:Evaluation` criterion under which the feature was judged (consistency). Adapted from Pasqual et al. (2024), Schimmenti et al. (2026).

and evidence and hypothesis extraction. The dataset includes 581 Wikipedia articles on historical forgeries, hoaxes, and authenticity debates across various categories. Of these, 45 articles were annotated in INCEpTION as GT, yielding 235 entities, 215 interpretation acts, 132 evidence nodes, and 308 Wikidata alignments. The evaluation used standard precision, recall, and F1 metrics, as well as G-EVAL (Liu et al. 2023b). The team tested three LLMs of different sizes, including Claude Sonnet 3.7 (Anthropic 2025), Llama 3.3 70B (Grattafiori et al. 2024), and GPT-4o-mini (OpenAI 2024). The evaluation was built on few-shot prompting and CoT reasoning. The pipeline operates at the paragraph level to avoid overlapping opinions between entities and reduce token usage. A surprising finding is that smaller models performed well, with GPT-4o-mini recognizing the most entities (77.3% of GT agents) despite having the fewest parameters.

ATR4CH and Viewsari share several technical similarities, but they differ in key design choices and goals. The main difference is that ATR4CH works with Wikipedia articles, which already state the scholarly claims that the graph will represent. In contrast, Viewsari uses Vasari's *The Lives* as a primary source, where interpretive claims about entities have not yet been clearly stated. This underscores the importance of the extraction process's provenance. For ATR4CH, claims can usually be traced back to a specific sentence in the source. For Viewsari, claims can only be checked by looking at the interpretive act that created them.

Commonalities include the use of XD for ontology engineering, INCEpTION for GT annotation (Klie et al. 2018), and PROV-O for provenance documentation (Lebo et al. 2013). However, their differences underscore the principal contribution of this dissertation. SEBI provides a detailed model of scholarly discourse concerning CH items. Rather than functioning as annotators that extract and annotate specific mentions, LLMs within SEBI extract scholarly agents, classify their authenticity judgments, identify the evidential features cited, attach evaluation criteria and

polarities, and extract alternative hypotheses. When the ATR4CH pipeline determines that a sentence constitutes a forgery judgment rather than a neutral observation or attributes an opinion to a named scholar rather than an anonymous source, each action represents an interpretive act performed under uncertainty, analogous to SEBI's modeling for Valla in Figure 5.6. However, while ATR4CH populates this representation using LLMs, in the resulting KG, their interpretive contributions are flattened, so, for example, Valla appears as the authoring scholar associated with his interpretation act, and the act is indistinguishable whether entered by a human cataloger or an LLM. Viewsari adopts a different approach, in which LLMs serve as the primary interpretive agent, with their outputs populating the KG. The extraction process, including its prompt, model version, reasoning steps, and the specific text used, is documented as structured provenance alongside the extracted claim. These approaches are complementary, addressing distinct aspects of the interpretive process. More on these steps follows in Chapter 7.

5.2.3.5 CIDOC CRM and large language models for Liao dynasty heritage

Wang and Zhang (2025) present a KG of CH from the Liao dynasty in Chifeng, constructed by combining CIDOC CRM with GPT-based extraction over different CH artifacts, including archaeological surveys and historical documents. The ontology is built on a minimized module of CIDOC CRM, consisting of `E27 Site`, `E53 Place`, `E52 Time-Span`, `E21 Person`, `E55 Type`, `E57 Material`, and `E7 Activity`, and extending it with Liao-specific subclasses for place names, architectural types, and inscription categories. Extraction proceeds through a CIDOC CRM-constrained prompt that specifies the entity and relation types, with predicates mapped onto CRM properties such as `P2 has type`, `P4 has time-span`, `P14 carried out by`, and `P53 has former or current location`. Table 5.1 reproduces the authors' comparison of human and GPT extractions on passages from *The Study of Song-Liao Diplomatic Relations*, in which the gap between reference annotation and model output is visible across five examples.

A complementary NER experiment on thirteen passages from the *Geographical Records of the Liao History* compares the generative approach to two supervised baselines and is reported in Table 5.2.

From the results, it is visible that generative models outperform rule-based and token-classification baselines on historical and out-of-domain text, particularly where surface forms are complex and training data scarce; the authors note correct recognition of compound temporal and geographical expressions such as "Shangjing Linhuang Prefecture," "Song Jingde Fourth Year," and "Tonghe Twenty-Fifth Year." What is most consequential for the present work, especially for the ObliquER pipeline, is the type of errors the authors encountered, which resemble those we encountered and report on in Chapter 9. This concerns, in particular, entity boundary errors, where the model

Table 5.1: Comparison of human and GPT extractions on passages from *The Study of Song-Liao Diplomatic Relations*, taken from Wang and Zhang (2025). Failure modes include silent omission, overgeneralization to a single type, temporal misreading, and correct extraction when the surface form is explicit.

Excerpt of paragraph	Human triplet example	GPT extraction example
In the seventh year of the Kaibao era of Emperor Taizu of the Song Dynasty... Prefect of Zhuozhou, Liao.	Zunyi Prefecture—P2—Ancient Ruins; Zunyi Prefecture—P4—Seven Years	No output
...Tomb of the Cavalry Commandant Teng Gong with inscriptions and Li Tao's military service and official duties...	Lusi Kongteng Public Cemetery—P2—Ancient Tombs; Lu Ming and Li Lin Tombstones—P2—Ancient Ruins	Both types of cultural relics are classified as ancient ruins
...In the battle of Gaoliang River, the Song army was heavily defeated. The Liao forces took advantage of the situation to advance and attack, while the Green Border Prefecture troops, due to their scarcity, had little respite.	Xi You Prefecture—P2—Ancient Ruins; Gaoliang River—P2—Mountain Water Land Names; Green Border Prefecture—P2—Ancient Ruins	No output
...The western attack on Daizhou, where Zhang Qixian was defeated... The two nations then resumed peace negotiations.	Xi Zheng Dai Prefecture—P2—Ancient Ruins; Xi Zheng Dai Prefecture—P4—Twenty-Seven Years	Misread the time as "Twenty-Seven Years"
...The northern march to Chanzhou, to strengthen the military force... Sent envoys with a letter to the Song governor of Mozhou, Shi Pu.	Zhu Zheng Yang City—P2—Ancient Ruins; Bei Xin Jiang Prefecture—P2—Ancient Ruins; Ren's Book Exemptions—P2—Ancient Ruins	Consistent with artificial results

incorrectly classified non-artifact phrases as artifacts, and missing recognition, where the model frequently failed to recognize all triples in sentences containing short place names.

Table 5.2: NER performance on thirteen passages from the *Geographical Records of the Liao History* (*Liaoshi · Dilizhi*), adapted from Wang and Zhang (2025). "Strict" and "Lenient@0.5" denote exact span matching and partial overlap at a 0.5 threshold, respectively; P, R, and F1 report precision, recall, and F1-score on the strict setting.

Method	Strict	Lenient@0.5	P	R	F1
`spaCy_zh`	0.300	0.375	0.214	0.268	0.250
`CKIP_bert_base_cn`	0.009	0.264	0.018	0.500	0.346
ChatGPT	0.452	0.590	0.410	0.642	0.505

The Wang and Zhang pipeline is valuable for what it reveals, especially the systematic errors it reports, such as entity boundary errors and missed recognition in passages with many short place names. These relate to the limitations we have elaborated on in Section 5.1.5.2, in which the idea is that the source text already contains clearly defined entities and that the extraction system's task is simply to recognize them. When this assumption is false, such as it is with implicit entity mentions that depend on historical context, performance drops in predictable ways. The KG

presents claims about Liao heritage but does not capture the reasoning steps that led to them, so prompts are not traceable; these are the key gaps that Viewsari aims to address, as discussed in Section 5.2.5.

5.2.3.6 FactGrid, a Wikibase for historical data

FactGrid is a collaborative research database built on Wikibase, the same open-source software used by Wikidata. It adapts this technology for academic historical research and has been part of the German National Research Data Infrastructure (NFDI) since 2023. FactGrid now has about 750 participants, 1.85 million pages, and over 19.25 million edits.[3]

It organizes information as items (Q-numbers) linked by properties (P-numbers) in a flexible data model. As an open repository, it allows researchers to add new properties as needed, and each statement can include qualifiers and source references. Researchers can create items for any entity they wish to study, even if it is obscure. This is especially helpful for historical entities missing from standard KBs, which would otherwise be considered OOKB entities. FactGrid supports many research projects, covering topics from ancient history to the twentieth century. Notable examples include:

- **Germania Sacra** (Göttingen Academy of Sciences): over 80,000 entries on ecclesiastical personnel of the Holy Roman Empire, representing one of the largest prosopographical datasets in the DH;
- **The Illuminati and Freemasonry Project**: the founding use case, modeling approximately 3,000 late-eighteenth-century individuals and their organizational networks.

FactGrid functions as the external reference point against which the Viewsari ontology is evaluated in the generalization study developed in Chapter 10.

We see FactGrid as an external anchor that allows the Viewsari ontology to show what it adds to existing KGs, with a dedicated provenance layer that separates human authority from computational extraction, while keeping both as first-class, citable records for the same passage of text (Figure 5.7; more in Chapter 10).

[3] FactGrid Wikibase: `https://database.factgrid.de/wiki/Main_Page`.

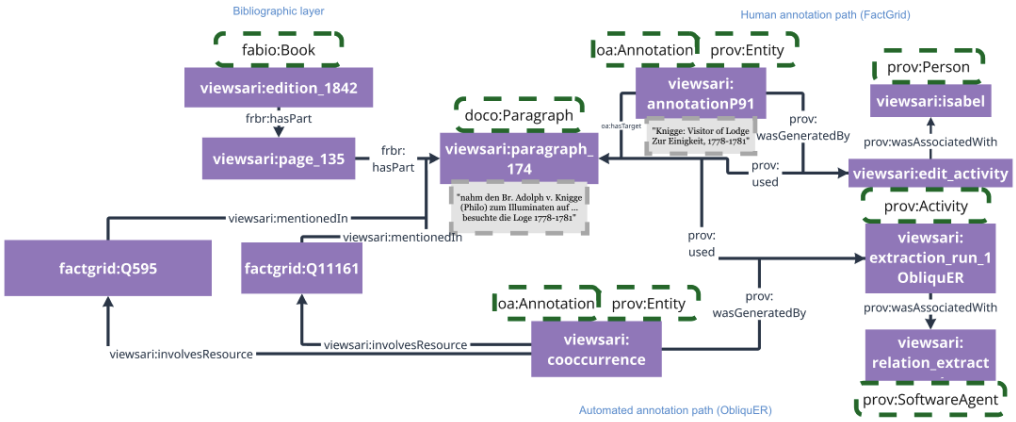


Figure 5.7: Dual-provenance representation of a co-occurrence in Viewsari: paragraph 174 of Kloß's *Annalen der Loge zur Einigkeit* (1842) serves as the shared `oa:hasTarget` for a human FactGrid annotation and an automatic ObliquER extraction, each recorded as a distinct `prov:Activity` but converging on the same aligned entities (*Knigge*, *Lodge Zur Einigkeit*). The Viewsari instantiation of this schematic on real GT data is shown in Figure 10.6.

5.2.4 Prior computational work on Giorgio Vasari's *The Lives*

There has been prior work on *The Lives* aimed at extracting historical knowledge and building a KG. Examining this work in detail is necessary both to situate the contributions of the present dissertation and to make explicit the representational limitations that motivate the methodology developed in Chapter 7.

5.2.4.1 Social network analysis based on name co-occurrences

The earliest approach to a quantitative, large-scale analysis of *The Lives* as quasi-structured data was produced by Santini et al. (2022). They constructed a network of artist co-occurrences for network analysis (based on the co-appearance of two-person references within a shared paragraph).

The extraction pipeline operates in two phases (Santini et al. 2024). Based on the Index of Names of the de Vere English translation, a web-scraper transforms the given output into a structured lookup table that maps artist names to the pages on which they are referenced and simultaneously segments the body of the text into paragraphs and assigns page numbers. In the second phase, a named entity recognizer based on LUKE (Yamada et al. 2020) identifies person mentions within each paragraph, and a coreference resolver (Otmazgin et al. 2022) clusters pronouns and name variants presumed to refer to the same individual. The Index of Names is then used to resolve

these clusters against the list of known persons. Built on these mentions, co-occurrences are formed when references to two artists appear within the same paragraph.

Co-occurrence strength was quantified using PMI, which assigns a positive score to pairs of artists who appear together more frequently than would be expected from their independent marginal frequencies, yielding a weighted edge in the resulting social network (Santini et al. 2024). The network released by Santini et al. (2024) comprises 479 weighted co-occurrence edges over 312 connected artists, drawn from the 1,061 persons of the de Vere Index of Names, and is published as a data deposit on Zenodo.[4] We re-examine and extend these statistics over our own extracted corpus, adding the Dice coefficient (Ondraszek et al. 2024b); the resulting co-occurrence analysis is detailed in Section 7.2.

5.2.4.2 ExtrART: an annotation reference for art historical information extraction

Santini et al. (2022) conducted experiments on both entity recognition and EL, including the development of a dataset and a set of annotation guidelines called ExtrART (Santini et al. 2023). Compiled from the same de Vere edition, it comprises 55 paragraphs from 27 chapters, annotated collaboratively using the INCEpTION platform (Klie et al. 2018). In a corresponding GitHub repository, a sample set of the following entity types is available: persons, organizations, places, and miscellaneous (all following CoNLL 2003 (Tjong Kim Sang and De Meulder 2003), partially linked to Wikidata, as well as artwork references, motifs, terms, and dates. For entity extraction, the authors employed UniNER (Zhou et al. 2024c).

The distribution of types in the result dataset is heavily skewed toward persons (`PER`: 732, 58.4%) and locations (`LOC`: 414, 33.0%), reflecting the biographical and geographical character of Vasari's prose. Organizations and miscellaneous references account for the remainder (4.5% and 4.1% respectively). Iconographic motifs (311 annotations) are linked to ICONCLASS (van de Waal 1974–1985) and divided into iconographic subjects (`SUBJ`: 281, 90.4%) and predicated states (`STATE`: 30, 9.6%); an additional 51 span-level entries make the *state-of* relation between the two types explicit. Artistic terminology (740 annotations) is linked to the Getty Art & Architecture Thesaurus (AAT) (Harpring 2010) and covers five sub-types: architectural objects (`OBJarch`: 279), generic objects (`OBJ`: 179), visual art objects (`OBJvis`: 149), materials (`MAT`: 76), and techniques (`TECH`: 57). Finally, temporal expressions (56 annotations) distinguish specific dates (`DATE`: 42, 75.0%) from vaguer time intervals (`INTERVAL`: 14, 25.0%), such as *after the death of Julius II*. Across all layers, ExtrART contains over 2,360 annotations.

[4] Santini's co-occurrence data and PMI tables: `https://zenodo.org/records/8395369` and `https://zenodo.org/records/8395425`.

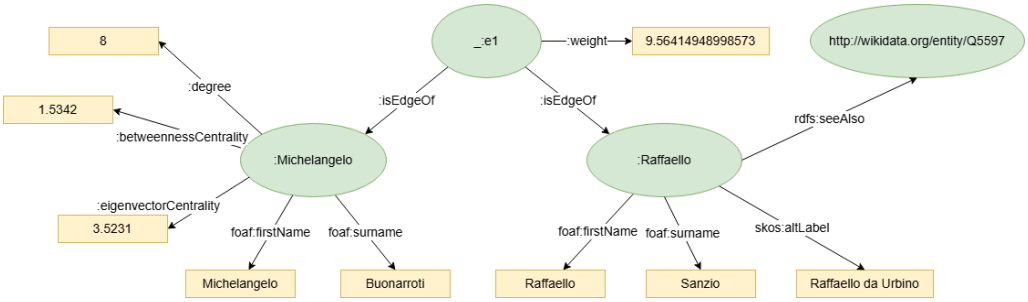


Figure 5.8: RDF reification of a weighted co-occurrence edge between Michelangelo and Raffaello in the social network, with the blank node `_:e1` carrying the edge weight and `:isEdgeOf` links to both participants; per-node centrality measures (degree, betweenness, eigenvector) are attached as datatype properties, alongside `foaf` name assertions and an `rdfs:seeAlso` alignment to Wikidata. Adapted from Santini et al. (2024)'s website.

As Figure 5.8 shows, additionally, the team has constructed an initial KG from this data, in which artists are represented as RDF resources and co-occurrences as weighted edges (Santini et al. 2022, 2024). The data model draws on standard vocabularies like RDFS and is published on Zenodo.[5] Each co-occurrence edge is represented as a blank node carrying a `:weight` property that encodes the PMI score, with `:isEdgeOf` properties linking the blank node to the two participating artist resources. Artist nodes are described using `foaf:firstName` and `foaf:surname`, `skos:altLabel` for alternative name forms, `rdfs:seeAlso` for external Wikidata identifiers, and a set of network centrality attributes: `:degree`, `:betweennessCentrality`, and `:eigenvectorCentrality` (computed from the full co-occurrence network). Per-volume network visualizations are additionally provided for each of the ten volumes of de Vere's translation.[6]

5.2.5 Limits of current approaches translated to requirements in Viewsari

Santini et al. (2022) provide a comprehensive reference annotation with ExtrART and demonstrate that general-purpose pipelines are inadequate for *The Lives*. Their work serves as a direct reference point for this dissertation, demonstrating the specific challenges the domain poses for implicit entity recognition and KG construction. Despite considerable contributions, there are gaps that come from the methodological choices made.

[5] Zenodo record for the preliminary KG: `https://doi.org/10.5281/zenodo.8395369`.
[6] Preliminary Vasari network visualization: `https://ise-fizkarlsruhe.github.io/vasari_network/`.

Both the social network and the ExtrART GT focus only on explicit entity mentions. As a result, many references are missed, especially implicit mentions and coreference chains that go beyond the immediate context in ExtrART. For instance, artworks mentioned indirectly in a biography, like "the gold-crested angel on the chapel ceiling" or "an angel crowning the chapel ceiling," are not recognized at all. For Viewsari and the ObliquER pipeline, we plan to improve the extraction process to identify implicitly mentioned entities as well. This effort includes developing a typology for implicit entities, which we discuss in Chapter 7.

In addition, the current KG uses only RDFS, so it cannot link extracted claims to their source text. Co-occurrence data lacks typed edges, making the meaning of these connections unclear, and there is no way to trace co-occurrences back to the specific paragraphs they come from. For example, a link between two artists is not connected to the text that establishes that link, even though a co-occurrence is essentially a direct result of mentions in a specific paragraph. Tracking provenance for human or machine annotation has always mattered, but it has often been overlooked because earlier methods appeared to offer implicit accountability. For example, attention weights indicate which words led to an entity tag; LIME (Ribeiro et al. 2016) explains classifier decisions; and human annotation assumes that reasoning can be traced if needed, even if it is not recorded in the data model. In reality, provenance was postponed rather than absent.

The challenge of provenance becomes acute when employing generative extraction methods such as UniversalNER within ExtrART. In contrast to fixed classifiers, prompt-based models are sensitive to parameters such as temperature, prompt formulation, and model version. Labels or co-occurrences produced by LLMs do not constitute fixed facts but rather probabilistic inferences conditioned by both the model and the prompt. What previously constituted a latent deficiency in data modeling now emerges as a manifest methodological problem. While generative artificial intelligence did not originate the provenance challenge, it renders its neglect untenable. Consequently, within the Viewsari framework, each extracted statement must explicitly record the interpretive process, including the prompt template, model version, and context window, as integral components of the knowledge graph.

To make this point clearer, many existing methods work well in fields with standard terms and stable entity identities. This is also true for manual interpretations of historical sources, such as in the HiCO project and the Vespasiano da Bisticci correspondence, where the HiCO ontology separates factual and interpretive statements, and the Musical Meetups ontology uses PROV-O to clarify extraction steps. However, most IE systems assume that entities are clearly named and that extracted content is factual. These assumptions work in areas such as news and movie ratings, but not in CH and historical narratives, where entities are often described rather than named, and extracted statements often require interpretation.

Two recent pipelines show this problem clearly. Wang and Zhang (2025), discussed in Section 5.2.3.5, demonstrate how this issue appears in practice. Errors in entity boundaries and missed recognition in ambiguous contexts occur when we assume that source texts contain clear, unambiguous entities ready for recognition, without considering broader context or implicit naming. Populating these results into a CIDOC CRM-based KG results in structures that collapse mention and referent without recording the source document; the chain of interpretive decisions disappears, and the triple customizes as a settled fact – precisely the move that historical source criticism exists to prevent. Schimmenti et al. (2026) focuses more on the epistemological side. Their ATR4CH method (see Section 5.2.3.4) combines generative extraction with the SEBI ontology, which uses RDF-star reification and PROV-O to represent different scholarly authenticity claims, including who makes them, the evidence they use, and the provenance of each interpretation. Provenance is tracked for who said what about the Donation of Constantine, but not for how the extraction pipeline made its choices. The errors reported by Schimmenti et al. (2026), such as false positives due to entity boundary mistakes, indicate unrecorded model-level interpretation. When these extraction decisions matter but are not documented, the KG lacks accountability that cannot be fixed later. Even within this epistemically sound framework, these issues become clear when examining the evaluation results. Metadata extraction gets F1 scores of 0.96–0.99 and evidence identification of 0.95–0.97, but hypothesis extraction drops to 0.65–0.75, and overall discourse representation is only 0.62 on G-EVAL. Extraction works well for factual metadata, but fails when the knowledge is constructed.

This gap is especially clear when LLMs are used for extraction. If we do not record prompts, model versions, context, confidence, or uncertainty, the results appear as fixed facts instead of hypotheses. Unlike earlier methods, the language model serves as both annotator and interpreter, making decisions based on large background datasets whose contents are often unknown (Zhou et al. 2024a). For Viewsari, this means the ontology should treat co-occurrences and implicit entity mentions as constructed objects within interpretive processes, each with its own epistemic status, rather than as undisputed facts. Currently, there is no approach that brings together interpretive epistemology, provenance-aware and reflective modeling for LLM extraction, and detailed entity representation within a single framework. As a result, the main contribution is not only improved extraction performance but also a new perspective that frames extraction as the creation of knowledge within semantic representations.

In this dissertation, we address this gap by adding representational expressivity, explicitly representing that information in the KG as resulting from specific activities under defined circumstances, including text anchorage and extraction provenance at the parametric level. Additionally, we address epistemic adequacy, considering whether the model not only allows but also requires these distinctions to be recorded, so that any representation created with the model is accountable by

design. The comparative modeling study in Section 9.2 examines both dimensions using CIDOC CRM, CRMdig, CRMinf, and the Viewsari model.

5.2.6 Knowledge graph visualization and interfaces for historical data

Building a KG alone does not make its information easily accessible to the scholars who need it most. Domain experts in the humanities, be it art history, musicology, or literary studies, rarely use SPARQL endpoints or RDF serializations directly. This has already been demonstrated by other projects, including the Sampo team from Finland, which released the Sampo User Interface (UI) to enable different ways to explore their diverse KGs (Ikkala et al. 2022). This need is especially important for historical and interpretive KGs, where questions are often answerable only ambiguously, such as "Who painted alongside Botticelli?' or "Which musicians met in Rome between 1600 and 1700?' Designing such interfaces is a research challenge that brings together information visualization, semantic web engineering, and DH methods.

Since the Viewsari project features a web interface as its main result (cf. Chapter 8), a comparison to other KG visualization approaches for historical data is necessary. One of the main features we want to integrate is fostering serendipitous findings (Waitelonis and Sack 2012), thereby supporting users in generating new hypotheses and ideas from the views they are offered; that is why we include examples with exploratory features for their KGs in the following sections. Furthermore, we review general-purpose graph visualization tools as an orientation baseline.

5.2.6.1 Domain-independent ontology and knowledge graph visualization approaches

Various tools for visualizing ontology and KG content exist. We introduce two of those, the dynamic WebVOWL application and the static notation standard Graffoo.

Interactive OWL visualizations on the Web. One of the first visualization tools is *WebVOWL* (Lohmann et al. 2016), which visualizes ontologies and instance data as a node-link diagram, enabling exploration of the graph structure itself. It is an established tool in this category that renders OWL ontologies as force-directed graphs directly in the browser. As Figure 5.9 illustrates, it provides a compact visual notation (VOWL) in which classes appear as circles, object properties as directed arcs, and datatype properties as rectangular nodes. While primarily designed for ontology communication and documentation, making it a built-in component of the

WIzard for DOCumenting Ontologies (WIDOCO) standard (Garijo 2017), WebVOWL has been widely adopted in DH projects as a lightweight means of communicating ontological structure to domain experts unfamiliar with OWL syntax.

Static ontology visualizations based on Graffoo. A complementary approach to ontology visualization is the Graphical Framework for OWL Ontologies (Graffoo), a diagrammatic notation developed by Falco et al. (2014) for rendering classes, object and data properties, restrictions, and axioms in a manner close to the OWL 2 semantics but legible to non-specialists. Unlike WebVOWL, Graffoo diagrams are manually authored and are not interactive. However, the notation has de facto become a standard for static diagrammatic conventions.

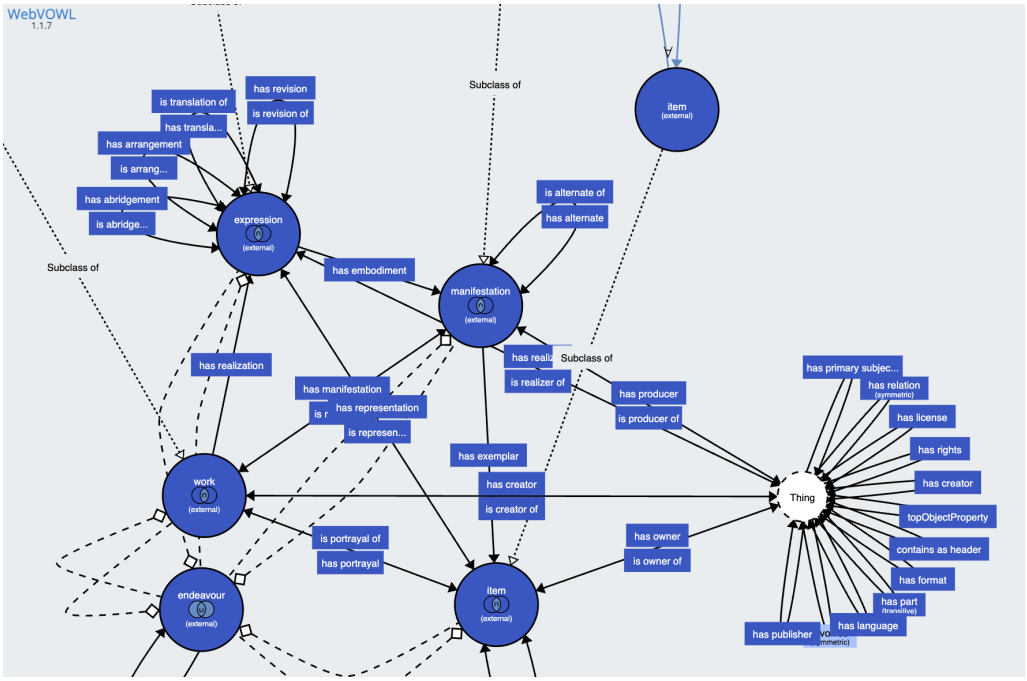


Figure 5.9: Excerpt from the WebVOWL visualization of the Viewsari ontology, generated via WIDOCO. Classes appear as blue circles and object properties as labeled arcs, with dashed lines indicating `rdfs:subClassOf` relationships to externally imported classes (here from FRBR and FaBiO). The excerpt shows the alignment between Viewsari's FRBR-derived core—`Work`, `Expression`, `Manifestation`, and `Item`—and the corresponding external classes.

For documenting the Viewsari ontology, we combine both paradigms. The WIDOCO-generated HTML documentation includes a WebVOWL rendering of the full ontology in a separate browser window. Alongside this, the Viewsari documentation portal embeds an interactive ontology graph

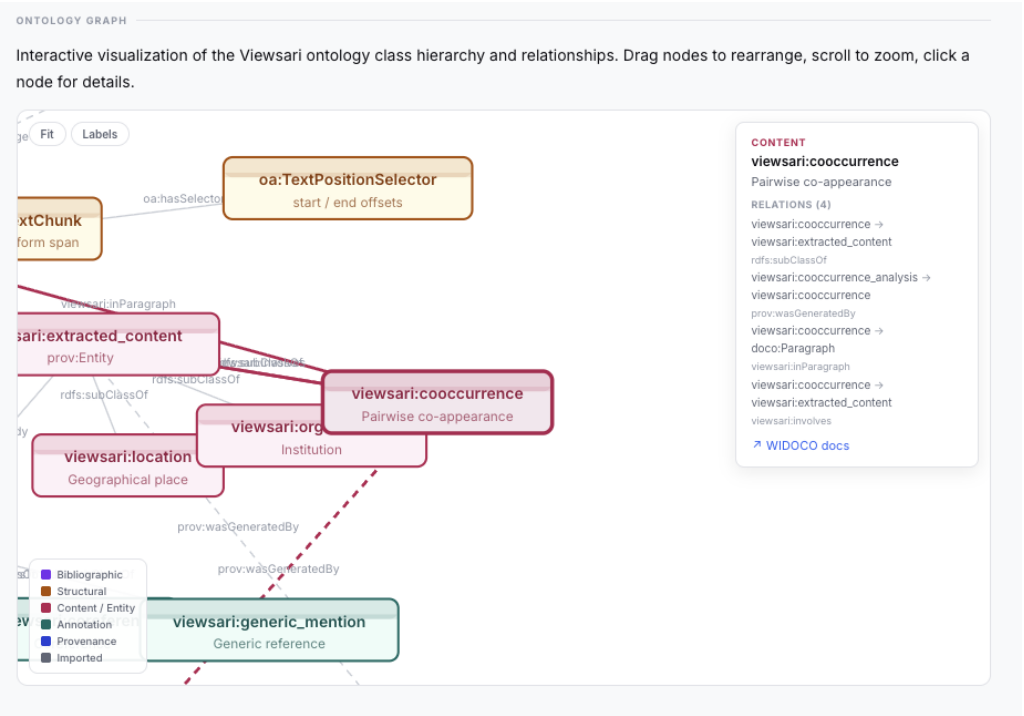


Figure 5.10: Interactive D3.js rendering of the Viewsari ontology as embedded in the documentation portal. Classes are grouped and color-coded by modeling module (bibliographic, structural, content/entity, annotation, provenance, imported), with a side panel exposing the outgoing and incoming relations of the selected class—here `viewsari:cooccurrence`—and direct links to the corresponding WIDOCO entries. Complementing the WebVOWL view that ships with the HTML documentation, this custom visualization foregrounds the ontology's modular, provenance-aware structure.

(cf. Figure 5.10) that highlights the modular structure and shows hierarchical structures with visual cues in different colors. For the ontology visualizations in this dissertation, we use an adaptation of the Graffoo framework, in which classes, instances, and other boxes receive different colors for easier separation and a more aesthetically pleasing appearance. More on this follows in Chapter 7.

Another common approach is to use SPARQL-powered faceted search and exploration interfaces. These systems let users explore a KG using filters, full-text search, and various result views, such as tables, maps, or timelines, all based on SPARQL queries. An example for this is given by Wikidata, where users can either execute SPARQL queries without additional features on the data via the Wikidata Query Service[7] (and output can further be formatted into tables, charts, an

[7] Query Service: `https://query.wikidata.org/`.

image grid, etc.), use visual helpers like Cita[8] for citations metadata or QueryGraph[9] to build a graph with nodes and edges to generate a query in SPARQL, or visualize data with Crotos[10] for visual artworks or Entitree[11] for a family tree of any human-like item in Wikidata.

Similar approaches can be implemented using JavaScript graph libraries such as D3.js.[12] These, exemplified by D3.js, are specifically designed to handle and visualize network data, enabling the translation of interconnected data points in a KG directly in the browser from JSON or RDF files. For example, Santini et al. (2022) used Pyvis[13] (a Python tool based on Vis.js[14]) to create co-occurrence network visualizations for the Vasari corpus.

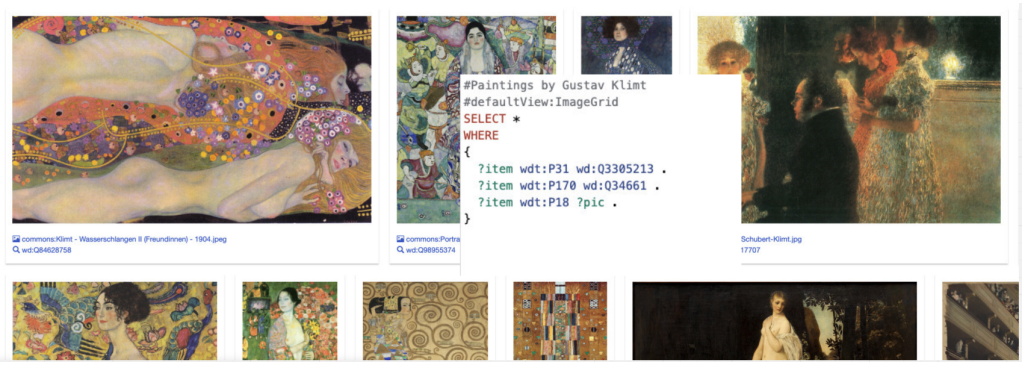


Figure 5.11: Example of a SPARQL-driven result visualization in the Wikidata Query Service. The query (overlaid) retrieves all instances of paintings (`wdt:P31 wd:Q3305213`) whose creator (`wdt:P170`) is Gustav Klimt (`wd:Q34661`), along with their associated images (`wdt:P18`). The `#defaultView:ImageGrid` directive instructs the interface to render the results as a grid of thumbnails rather than a tabular listing, illustrating how lightweight view hints embedded in SPARQL queries can yield domain-appropriate presentations of CH data without bespoke interface code.

5.2.6.2 Digital humanities and art history-specific visualization tools and interfaces

In the DH, the specific needs of historical and CH data, along with related research queries, have led diverse projects to develop specialized visualization tools. These tools often treat the graph as a basis for exploring and interpreting connections within structured collections (Rantala et al.

8 Cita: `https://www.wikidata.org/wiki/Wikidata:Zotero/Cita`.
9 QueryGraph: `https://datavizdev.fr/QueryGraph/index.html?lang=fr`.
10 Crotos: `https://zone47.com/crotos/`.
11 Entitree: `https://www.wikidata.org/wiki/Q104728828`.
12 D3.js: `https://d3js.org/`.
13 PyVis: `https://pyvis.readthedocs.io/en/latest/`.
14 Vis.js: `https://visjs.org/`.

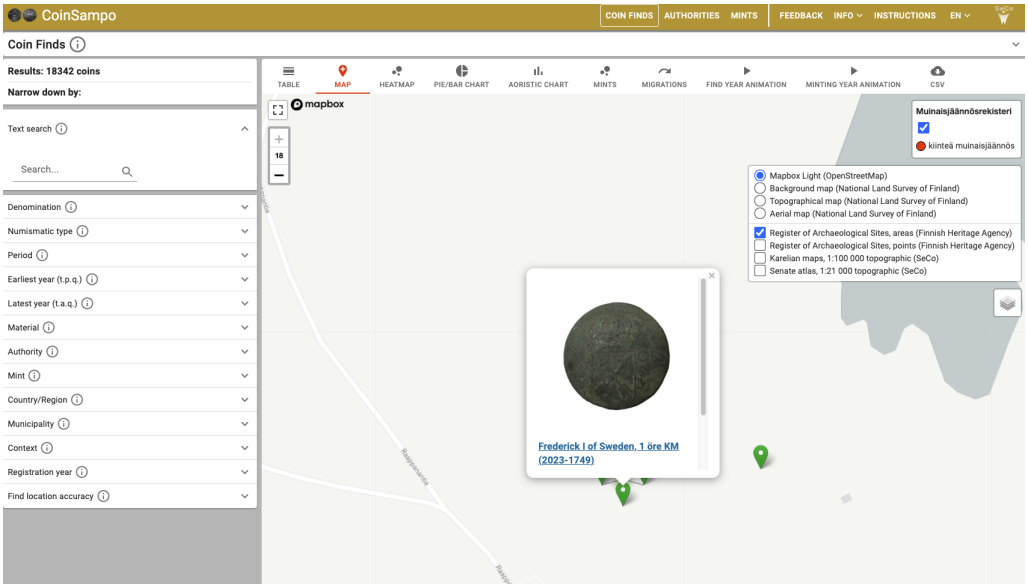


Figure 5.12: The *CoinSampo* portal, built with the Sampo-UI framework, presents a linked-data graph of Finnish numismatic finds. The left panel provides faceted filters for denomination, numismatic type, period, earliest/latest years, material, authority, mint, and findspot. The right panel offers interchangeable perspective tabs, such as table, map, heatmap, pie/bar chart, aoristic chart, migrations, minting-year animation, and export, all applied to the same filtered result set. The map displays the Finnish Heritage Agency's archaeological site polygons on an OpenStreetMap basemap, with individual coin finds shown as markers. Selecting a marker opens an entity card (Rantala et al. 2024b)

2024a). We present two of those endeavors in the following subsections, based on the Sampo KGs that we briefly mentioned in Section 5.2.3 and on the demo of the In/Tangible European Heritage (InTaVia) project (Ebel et al. 2023, Windhager et al. 2024).[15]

The Sampo model and Sampo-UI. The Sampo model (Hyvönen 2023, Ikkala et al. 2022) provides a systematic, well-documented approach to designing KG interfaces for CH data. This framework has been implemented in a wide range of Finnish and international CH projects, such as *CoinSampo*, an interface for Finnish numismatic finds developed by the Finnish Heritage Agency (cf. Figure 5.12) (Ikkala et al. 2022, Rantala et al. 2024b).

Each Sampo portal is organized around three canonical view types, namely a landing page that provides entry points into one or more faceted-search perspectives; perspective views, each centered on a specific entity class and combining facets with multiple result-set visualizations;

[15] InTaVia project page: `https://intavia.eu`.

and entity landing pages that present detailed metadata for a single resource, including internal and external recommendation links.

From a design perspective, the Sampo model is facet-driven rather than query-driven, thereby reducing barriers for domain experts. It offers diverse perspectives on historical artifacts, as seen in the coin example, where the tab provides options to explore the coin's context (see Figure 5.12 for example, the possibility to switch between a map heatmap, a table, etc.). These characteristics position the Sampo framework as a direct reference point for the Viewsari interface, which also aims to enable users to explore the data from different perspectives.

Visualiziations of interrelated biographical knowledge in InTaVia. The InTaVia project is a Horizon 2020 research and innovation initiative that integrates national biographical databases from across Europe. As one core deliverable, the project aims to represent historical KGs in accessible visualizations and to demonstrate how to represent multiple dimensions of cultural data within a single coherent visual framework (Fokkens et al. 2018, Windhager et al. 2024).

Their KG of in/tangible CH encompasses historical actors, events, places, and cultural objects that are interconnected entities (Windhager et al. 2024). The InTaVia interface (cf. Figure 5.13) provides entity landing pages that display a person's connections as an interactive ego-network alongside a chronological timeline of their creative output, thereby integrating biographical and cultural-object data. Architecturally, InTaVia is similar to Sampo, employing a SPARQL backend and building perspective views on top of it.

For Viewsari, the primary relevance lies not in the specific technique but in the broader methodological commitment to integrating diverse perspectives on biographical data, including not only information about persons but also related context from their biographies, such as artworks, other documents, and places. The shared emphasis is on the hypothesis that interpretive humanities KGs cannot be reduced to a single visual perspective and require interfaces that integrate multiple epistemic dimensions.



Figure 5.13: Entity landing page for the Belgian history painter Louis Gallait in the InTaVia visual analytics studio. The upper panel renders an ego-network of entities connected to Gallait (CH objects as green squares, events as green circles, places as purple triangles), while the lower panel arranges his attested events and productions along a horizontal timeline (c. 1835–1875). The interface exposes external authority links (Wikidata, GND, VIAF) and situates the actor within the InTaVia integrated KG of in/tangible European CH (Windhager et al. 2024). Screenshot taken from `https://intavia.acdh-dev.oeaw.ac.at/`.

6 “As long as art lives never shall I accept that men are truly dead:” Giorgio Vasari’s *The Lives*

In the preceding chapters, we have established central claims for this dissertation’s argument. Existing approaches to KE and KG construction, namely advanced generative extraction systems, remain oriented toward detecting explicit surface forms and normalizing entities against existing KBs. Yet humanities sources fundamentally oppose these assumptions, as they encode meaning through different forms, such as implicit mentions, which influence how the extraction process itself must be modeled. When ambiguity is either forcibly flattened through disambiguation or dismissed as noise, long-tail and OOKB entities persist as barriers rather than as means for better understanding. Critically, such pipelines rarely account for the epistemic status of extracted statements or the provenance of interpretive decisions.

If KGs are meant to capture interpretation rather than just record facts, then the way we evaluate them needs to change as well. A good case study should test the basic ideas behind traditional extraction methods and see if an approach that tracks sources and interpretation can handle ambiguity without oversimplifying it.

This chapter presents our case study corpus, created to meet these needs. By formally introducing Giorgio Vasari’s *Le Vite de’ Più Eccellenti Pittori, Scultori, e Architettori* (*The Lives*) (Figure 6.1) (Vasari 1568), we set out the criteria an evaluation corpus should meet for our research and explain how we built the raw corpus from Vasari’s work. We also show why it is a good test case. The chapter ends by describing the data and annotation methods used for KE. Details about building and sampling the annotated GT corpus are given in Section 7.4.1.

6.1 Advantages of a case study approach

To fully ground this dissertation’s representational commitments, we want to elaborate on the reasoning behind using a case study approach and, specifically, why *this* use case. The decision

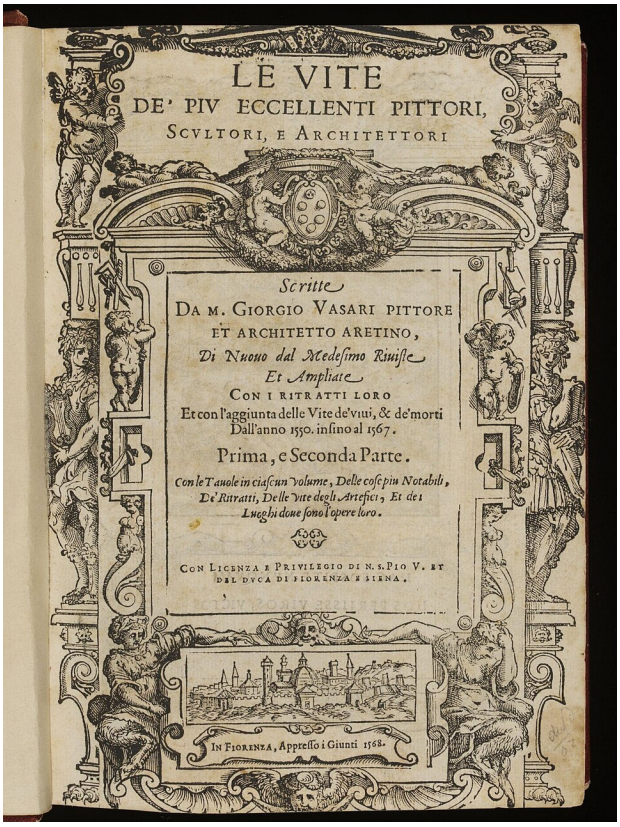


Figure 6.1: Frontispiece of the second edition of Giorgio Vasari’s *Le Vite de’ Più Eccellenti Pittori, Scultori, e Architettori*, published in Florence by the Giunti press in 1568. The 1568 edition is the source text for the Viewsari corpus. Available via Wikimedia at `https://commons.wikimedia.org/wiki/File:Vite.jpg`.

is deliberate, reflecting a research design choice that aligns with case-based inquiry across the social sciences, software engineering, and the DH.

A case study, in the sense developed by Crowe et al. (2011), generates an in-depth, multi-faceted understanding of complex issues in real-life contexts. This approach suits questions of *how* and *why* where boundaries are not cleanly separable from context, and situations in which the goal is to understand interpretive mechanisms rather than to estimate, i.e., the accuracy of a system from a single perspective. Thus, we designed the three RQs outlined in Chapter 1 to align with this approach. Formulated in the tradition of design science research (Hevner et al. 2004, Wieringa 2014a), these RQs treat a constructed artifact and its justification as the unit of inquiry, rather than a population parameter. The appropriate evidence is a working ontology, a documented pipeline,

a populated KG, and a transferability demonstration, each of which constitutes an answer to a *how* question, not an estimate of a quantity.

Wieringa (2014c,d) articulates this distinction as an investigation of architectures and mechanisms of individual systems and generalization by analogy. Vasari's *The Lives*, thus, can be treated as a system of such components, manifesting an architectural object in the same sense. The mechanisms by which Vasari constructs interpretation, by which an LLM-based pipeline recognizes implicit entities, and by which provenance is documented or lost are case-based phenomena. They are studied here in detail in one case, with the explicit intention of producing analogic generalizations to other interpretive corpora rather than statistical generalizations to a defined population of texts.

Within the typology proposed by Crowe et al. (2011), the present study combines mostly intrinsic and instrumental motivations. It is intrinsic in that *The Lives* is of genuine interest in itself, owing to its canonical status in art history, its layered publication history, and its dense interweaving of explicit and implicit references, which makes it a phenomenon worth understanding on its own terms. It is instrumental in that the case also helps develop a broader argument about generative KE, particularly the PIEA and the provenance gap, which are transferable well beyond Vasari. This dual character is consistent with how case studies in the DH are typically motivated. Olieman et al. (2017) make a similar argument in their work on EL for parliamentary discourse, where the Dutch Golden Age and World War II are selected not because they represent all historical events, but because each exemplifies a form of interpretive complexity that exposes the strengths and limitations of the examined systems. *The case, like Vasari's work, is chosen because it is difficult, not because it is average.*

The Lives meets several criteria for a strong case study corpus in DH research, following Crowe et al. (2011), who emphasize the importance of a clear boundary that defines the scope, the evidence to be collected, and the main priorities for analysis. In this dissertation, the corpus boundary is clearly defined: it includes only the 1568 edition, uses de Vere's 1912 English translation as digitized by Project Gutenberg, is divided into paragraphs, and is limited to 168 biographies across ten volumes (cf. Section 6.4.2). The choice to use this edition and translation, instead of the original Italian or other English versions, is based on accessibility, consistent formatting, and the availability of tools for modern English. Vasari's writing is especially valuable for this case study because it reveals complexities that simpler texts would not. The text includes explicit, coreferential, and implicit mentions of entities that interact closely, highlighting the limits of extraction methods based solely on surface forms. Since it exists in several editions and translations, it also shows the need for provenance-aware representation. Finally, because the text is written as an interpretation rather than a factual report, it demonstrates why extracted

statements cannot always be treated as facts about an unchanging world. This allows both the technology and the theory to be tested in meaningful conditions (Olieman et al. 2017).

A common criticism of case studies is that they do not allow for generalization (Crowe et al. 2011). However, this concern mainly relates to statistical generalization, not to analytic or analogical generalization, in which mechanisms found in one case may also apply to similar cases. The methodology presented here, including the Viewsari ontology, the ObliquER pipeline, and the approach of treating generative extraction as an interpretive process requiring documentation, is intended to be reused. Chapter 7 revisits this idea by highlighting recurring design patterns in provenance-aware KGs in the DH and by discussing when the Viewsari methodology can be adapted to other interpretive sources. Here, transferability follows Crowe et al. (2011): it does not mean that findings will apply to every text, but rather that the analytic steps are sufficiently explicit for other researchers to judge whether they fit their own work.

6.2 Requirements for a case study corpus

Choosing a corpus in DH is almost never a neutral step. As Bogetić et al. (2022) point out, decisions about the source, scope, segmentation, and annotation all influence which questions can be asked and which answers are possible. This means that compiling the corpus is part of the research itself. The same is true in this study. The selection of *The Lives*, the use of paragraph-level segmentation, the sampled annotation corpus, and the handling of OOKB entities, choices described in Chapter 7, are not just technical steps. They put our central claim into practice, showing that extracting knowledge from interpretive texts is a constructive process. Choosing a different case, or even approaching this one differently, would lead to a different thesis. The case study does not just illustrate the argument; it is the argument in empirical form.

As mentioned in the introduction to this chapter, most current extraction and KG methods look for clear, explicit references. To test the interpretive modeling approach, the corpus needs to go further and include implicit references, in which entities are described rather than named. To meet these requirements, we chose Giorgio Vasari’s *The Lives* as our corpus. We will describe it in more detail in the next section.

6.3 Publication history and details about *The Lives*

Giorgio Vasari’s *The Lives* is both historically influential and structurally complex. First published in Florence in 1550, it is seen as one of the founding texts of art history and as the first to bring

together artist biographies in a new way. Vasari's stories about earlier and contemporary artists, from Cimabue to Michelangelo, shaped how later artistic periods and the Renaissance were understood by scholars and laypeople alike, and his work remains a standard part of art-historical education.

Originally, Vasari's Italian version of *The Lives*, published in 1550 (and revised in 1568), consists of three volumes, following the chronological order of artistic development during the Italian Renaissance. The first stage begins with artists such as Cimabue and Giotto, whom Vasari credits with reviving art as a naturalistic representation after what he describes as a decline in the quality of medieval art. In the second stage, Vasari retells stories of artists such as Donatello and Filippo Brunelleschi, whose work in perspective, anatomy, and architecture helped lay the foundation for Renaissance art. Finally, the third volume encompasses his reconstruction of the 'peak' of Italian Renaissance art, with artists like Leonardo da Vinci, Raphael, and especially Michelangelo, whose art he describes as the highest form of artistic achievement.

Regardless of his claim to objectivity, his biographies are not neutral accounts of the artists' lives. Vasari shares anecdotes about their training, travels, patrons, and professional relationships, and describes their major works and the development of their styles. He also includes presumptions about artists' personalities or creative talents, with hints at artistic rivalries, giving a detailed look not only at what artists produced at that time but also at how they were connected and the social world of the time, as seen in whom he chose to highlight. He also interprets artworks in terms of stylistic progress, showing how each artist contributed to the development of Renaissance art.

Consequently, this textual framework creates a complex web of connections among these entities, including artists, artworks, cities, and historical events. These links can often be explicitly stated, with relations among artists, such as student-teacher links, or artworks as outcomes of specific artists' processes. Vasari presents this interconnected world from his own perspective, as someone involved in it.

From a publication-history perspective, the canonicity of *The Lives* becomes evident in how it has been received and altered over time. The first Italian edition, published in 1550, included a substantial collection of biographies organized by artistic progression. In 1568, Vasari published a substantially revised and expanded second edition, incorporating numerous new biographies and updating earlier ones. Because of these additions, the 1568 edition is widely considered the standard version of *The Lives* in modern scholarship (Pon 2014). His work continues to influence art history; as Lisa Pon states, "since its first publication in 1550, those seeking to learn about Italian Renaissance artists have read Vasari's Lives of the Artists" (Pon 2014). Burioni (2017) identifies Vasari's work as a foundation for art theory and, by extension, for art history as a discipline, and notes the myriad artist biographies that have shaped the Renaissance as a historical concept.

Art-historical research has identified multiple variants of the text and analyzed its content, confirming some biographical details while refuting others. This work has resulted in several critical historical editions, including the German Wagenbach edition (Nova 2004) and English translations by Jonathan Foster (Vasari 1850/1852) and Gaston du C. de Vere (Vasari 2008). Despite extensive scholarly efforts on *The Lives*, art history lacks a structural representation of their content that systematically prepares the biographies for extended inquiries such as network analysis or the analysis of organizational affiliations. Stephen J. Campbell proposed a research program to interpret Vasari’s text as a comprehensive system of diverse and competing positions, advocating for a concept of “historical implicatedness” (Campbell 2008). Viewsari is designed to advance this understanding by improving access to and discovery of the text’s multiple perspectives and by unlocking new dimensions of interpretation.

As outlined in Chapter 1, this dissertation investigates three RQs, and Vasari’s *The Lives* offers a test bed in which all three can be operationalized:

- RQ1: What representational commitments must a KG make to remain accountable to the interpretive act that produced it?
  - *The Lives* is not neutral; it constitutes a view through Vasari’s lens on aspects of the Italian Renaissance. Thus, statements are epistemic qualifications, and artworks are embedded within subjective narratives of artistic progress. Modeling this material requires representing different ways entities are mentioned (implicitly, explicitly, ...?) and framed by interpretive acts.
- RQ2: Under what conditions can LLMs recognize and link the implicit and OOKB mentions that surface form-baselines miss?
  - Entities in Vasari’s work are frequently implicitly rather than explicitly named. Artworks may be identified by iconographic description (e.g., “a panel representing the Assumption”), by location (e.g., “in the chapel of Santa Croce”), or by additional context. These implicit references challenge surface-form-based NER and conventional EL pipelines. The corpus, therefore, provides a testbed for evaluating generative, prompt-based recognition of implicit entities and long-tail linking.
- RQ3: What reusable ontology design pattern emerges from modeling extraction as a documented, provenance-aware interpretive process, and *how* does that foundation transfer to structurally distinct (interpretive) corpora?

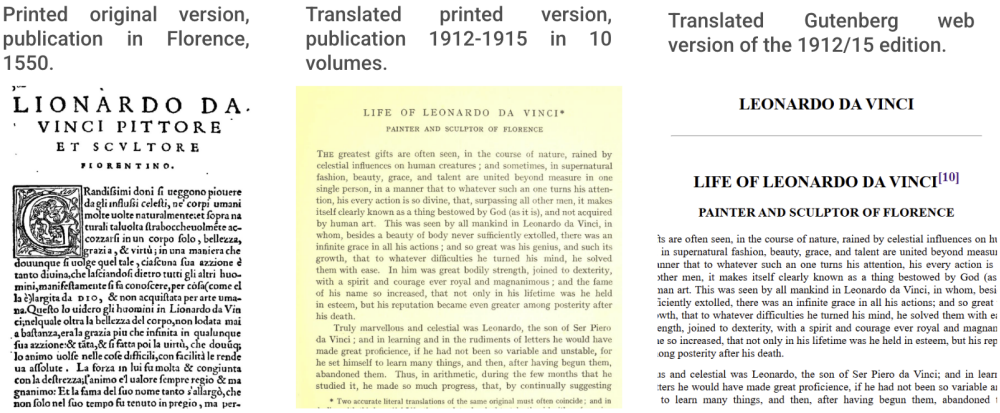


Figure 6.2: Publication history of Vasari's biography of Leonardo da Vinci, showing the 1550 original Italian printing (left), the 1912 English translation (center), and the modern web-based digital version (right).

– *The Lives* exists in multiple editions (1550, 1568) and several translations (see Figure 6.2), accompanied by centuries of reworks, comments, and the like. This multiplicity introduces variation in wording, attribution, and interpretive emphasis. Modeling such variation requires explicit transparency and a provenance-aware representation, which are essential features of a transferable methodology applicable to other interpretive, historical, or multilingual corpora. Similarly, the recognition of implicit entities plays a key role in other domains, including scientific publications and sentiment analysis (Schimmenti et al. 2025).

6.4 Data foundation: sources and pre-processing

In the following subsection, we describe the construction of a dataset that meets the aforementioned criteria, based on the English translation of Giorgio Vasari's *The Lives*. This description also includes text extraction and preprocessing.

6.4.1 Corpus and dataset construction from the Gutenberg project

This project utilizes the English translation by Gaston du C. de Vere (1912), based on the 1568 edition (Vasari 2008). A digital edition of this translation is available in ten volumes through

Project Gutenberg.[1] De Vere’s translation spans ten volumes, compared to the three volumes of the original Italian edition, owing to the expansion of the text through translation and the inclusion of supplementary editorial material. It follows the expanded second edition. The corpus contains 168 artist biographies, along with introductory material and supplementary indices, such as an Index of Names, which is shown in Table 6.1 and included in the translation.[2]

Table 6.1: Example of an Index of Names, an excerpt of the Index of the Craftsmen Mentioned in Volume II of the de Vere translation (Project Gutenberg, `https://www.gutenberg.org/files/25759/25759-h/25759-h.htm#Page_273`).

Name	Page(s)
Agnolo Gaddi	15, 25
Andrea della Robbia	125–127, 175
Andrea Pisano	50, 81, 83, 91, 93, 120, 145, 147, 154, 160, 200
Angelico, Fra (Fra Giovanni da Fiesole)	190, 271
Angeli, Don Lorenzo degli (Don Lorenzo Monaco)	55–58, 171

This English translation was selected because it is accessible and consistently formatted as a digital version of the text, and because English allows the use of widely available NLP tools optimized for modern English. While the original Italian text would offer philological advantages, we lack the resources to develop a robust tool for early modern Italian. Finally, the de Vere translation, even in digital form, preserves the structural organization of the printed volumes, including chapter divisions and indices, thereby facilitating future alignment between the digital text and the historical source.

6.4.2 First statistical overview of the Viewsari corpus

After following the extraction procedure described above, the Viewsari corpus includes 168 individual biographies spread across the ten volumes of the de Vere translation. In total, there are 3,026 pages and 3,479 paragraphs. Table 6.2 provides a summary of these structural details.

[1] Gutenberg project homepage: `https://www.gutenberg.org/` and for an overview of the volumes, see the Onlinebooks list `https://onlinebooks.library.upenn.edu/webbin/metabook?id=livespainters`.

[2] An Index of Names is a reference section in a book that lists all persons mentioned in the text, usually arranged alphabetically, along with the pages on which they appear.

The biographies are spread unevenly across the ten volumes. Volume 3, which covers the second 'age' (artists like Donatello and Brunelleschi), has the most biographies, with 33. In contrast, Volumes 9 and 10 each have only 8 biographies. This uneven distribution shows a focus on the High Renaissance and the addition of supplementary lives in the later volumes of the translation.

The biographies themselves vary considerably in length. The longest biography is Michelangelo Buonarroti's, which spans 150 pages and 227 paragraphs. On average, a biography contains 18 pages and 21 paragraphs, with a median of 11 pages and 14 paragraphs, indicating a long-tailed distribution dominated by a small number of substantially longer lives. Measured in tokens, an average biography comprises approximately 5,300 words (median: 3,200), with a minimum of 167 and a maximum of 57,566.

At the paragraph level, the unit chosen as the basic segmentation granularity for annotation has a mean length of 1,396 characters (median: 1,171) and 251 words (median: 211). The average word length across the corpus is 4.57 characters, consistent with general English-language prose. In total, the corpus contains roughly 4.86 million characters and 872,000 word tokens.[3]

Table 6.2: Structural statistics of the Viewsari corpus, computed from the extracted de Vere translation of *The Lives*.

Statistic	**Value**
Volumes	10
Biographies	168
Pages	3,026
Paragraphs	3,479
Word tokens	871,616
Characters	4,857,607
Avg. pages per biography	18.3
Avg. paragraphs per biography	21.2
Avg. words per biography	5,315
Avg. paragraph length (words)	251
Avg. paragraph length (characters)	1,396
Avg. word length (characters)	4.57

[3] In Table 6.2, the *Biographies* count of 168 is the full source-corpus scope, whereas the per-biography averages are divided by the 165 biography instances materialized under `viewsari:biography`; the 3-unit gap is the supplementary entries detailed in Table 8.4.

Cross-page paragraphs. A notable structural feature of the de Vere translation is the high incidence of paragraphs that span more than one printed page. Of the 3,479 paragraphs in the corpus, 1,676 – 48.2 % – begin on one page and end on another, with the per-volume rate running from 42.4 % in volume 5 to 58.6 % in volume 2 (Table 6.3). This reflects Vasari’s prose style, which favors long, syntactically extended sentences and paragraphs that follow the rhythm of the biographical narrative rather than the boundaries of the printed page. The phenomenon has direct consequences for the ontology design, motivating the introduction of a `viewsari:has start page` / `viewsari:has end page` property pair in place of a single-page provenance link (Section 7.3.2.3).

Table 6.3: Cross-page paragraphs per volume in the de Vere translation. A paragraph is counted as cross-page if its text body extends across at least two printed pages of the digital edition.

Volume	Paragraphs	Cross-page	%
1	287	122	42.5
2	251	147	58.6
3	343	151	44.0
4	392	177	45.2
5	375	159	42.4
6	410	205	50.0
7	355	195	54.9
8	383	184	48.0
9	379	196	51.7
10	304	140	46.1
Total	**3,479**	**1,676**	**48.2**

6.4.3 Text extraction and preprocessing of HTML content

Following Santini et al. (2022), the text was extracted paragraph by paragraph from its digital representation, preserving structural correspondence with the printed volumes. Paragraph-level segmentation was selected because it aligns with common digital editing guidelines, such as those recommended by the TEI (Sperberg-McQueen and Burnard 1994), which consider paragraphs as meaningful units for annotation and interoperability. This approach also ensures compatibility with existing text encoding standards and facilitates reuse in other DH projects.

The texts were scraped from the Gutenberg website.[4] According to the HTML representation's structure, the text was harvested paragraph by paragraph. These paragraphs match the printed original book's actual representation and were stored accordingly: To obtain a machine-readable version of the corpus, the text was automatically extracted from the digital Gutenberg edition using Python scripts. The HTML structure of the online volumes was parsed to extract the biographies' text, excluding navigational elements and other metadata present on the web pages. Each extracted paragraph was assigned two identifiers derived from the HTML structure: a page identifier corresponding to the original page numbering of the digital edition, and a paragraph identifier indicating its position on the page. This procedure preserves the structural correspondence between the digital corpus and the printed volumes while enabling fine-grained annotation and reference during extraction. The resulting dataset, therefore, consists of structured entries containing the page number, paragraph index, and normalized text content, as illustrated in the following text snippet:

```
page, paragraph, text
3, 1, "While the most noble and ..."
4, 1, "and arts-namely, in painting, in sculpture, ..."
4, 2, "There was born a son, then, ..."
...
```

Chapter 7 provides a detailed description of the further construction of two datasets, which are based on the extracted data from Vasari's texts and aim to facilitate analysis of both explicit and implicit entity mentions within art historical narratives and permit testing of the proposed approach across diverse textual environments and varying levels of referential explicitness. The aforementioned section outlines annotation guidelines and the preprocessing pipeline, which are central methodological contributions of the dissertation.

[4] Extracted content directory on GitHub: `https://github.com/ISE-FIZKarlsruhe/viewsari/tree/dev/src/extract_content`.

7 Methodology

In this chapter, we outline the methodological framework for the dissertation's research objectives. We begin with Section 7.1, which provides an overview of the proposed approach and its main components. Subsequently, the sections explain each methodological step in detail, including ontology engineering, IE, and the development of the ObliquER pipeline, KG population, and web application. Together, these comprise the full Viewsari framework.[1]

7.1 Proposed approach

As the central epistemological commitment, our claim that knowledge derived from interpretive sources must be accompanied by documentation of the conditions under which it was produced drives this dissertation's methodology. Operationalized as interpretation provenance, we attribute that, for every extracted entity mention, the Viewsari KG records (1) the source text span, (2) the extracting agent's parameters via a dedicated LLM agent class, (3) the prompt template, and (4) the temporal context of extraction. This does not reconstruct the model's internal inference; it documents the conditions under which that inference was elicited. The result is a KG in which every triple carries a complex provenance chain traceable to a specific sentence and a specific computational extraction process – the minimum necessary condition for *epistemically accountable* KE from interpretive texts.

To address the research objectives, the methodology translates the conceptual contributions **C1**–**C4** into four empirical artifacts **E1**–**E4**. Section 7.2 first constitutes a preliminary characterization of *The Lives* through co-occurrence analysis and a baseline assessment of entity type coverage, inspired by the work of Santini et al. (2024). Section 7.3 then develops **E1**, the Viewsari ontology, which operationalizes **C1** (the rejection of the PIEA and its three-layer representational response), **C2** (the provenance gap in generative IE) at the representational level, and **C3** (hermeneutic

[1] Parts of this chapter have appeared in preliminary form: the ontology engineering process, user stories, and competency questions in Ondraszek et al. (2024a); the corpus-level co-occurrence and network analysis in Ondraszek et al. (2024b); and the ObliquER pipeline in the ObliquER paper (under review, CIKM 2026). This chapter substantially extends, formalizes, and integrates these results.

provenance as conceptually distinct from XAI). The KE process using the ObliquER pipeline (**E2**) is described in Section 7.4; it instantiates **C2** at the system level by recording every extraction step as activity in the resulting graph. The formal task definition is given in Section 7.4.3, and the annotation strategy and GT construction in Section 7.4.4. Prompt engineering for entity recognition – covering the recognition pipeline architecture, prompt variants, dynamic chunking, and CoT prompting – is detailed in Section 7.4.8, and EL in Section 7.4.9. The pipeline is evaluated on a contemporary reference dataset in Section 7.4.10. Chapter 10 presents **E4**, the transferability pilot on 19th-century German documentary sources, which demonstrate that the design commitments of **C4** (Computational Provenance as a reusable ODP) are not Vasari-specific. All results are integrated into **E3**, the Viewsari KG; the population process is detailed in Section 7.5. Section 7.6 describes the deployment layer for the populated graph, including a documented ruleset, a SPARQL endpoint, and the competency-question evaluation framework that operationalizes the requirements catalog against that endpoint. All results are accessible via a web application, the design of which is discussed in Section 7.7. The abstraction of the methodology developed here into domain-neutral design patterns, thereby completing the realization of **C4**, is treated alongside another transferability pilot in Chapter 10 (see Section 10.3.2 and Section 10.3.3).

7.2 Corpus overview and baseline analysis

This section provides a *preliminary*, data-driven characterization of *The Lives* prior to ontology development. In Chapter 6, we explored why Vasari's work is suitable as a corpus and the characteristics that led to this conclusion. The initial co-occurrence analysis, initially conducted by Santini et al. (2022, 2024), has provided quantitative insights into the corpus structure and content distribution and has also informed subsequent decisions regarding ontological modeling and content extraction. A substantial extension follows in Section 7.4, once ObliquER has been applied to the corpus.

7.2.1 Understanding Vasari: Co-occurrence statistics and person coverage

As a first step toward the global structure of Vasari's work, we extended the co-occurrence analysis over the extracted paragraphs (Ondraszek et al. 2024b). It gives a corpus-level picture of the relational landscape of *The Lives* – which artists are discussed together, how that distribution is structured, and which figures are most canonically prominent. To do so, we re-examined the analysis introduced by Santini et al. (2024), whose pipeline is described in Section 5.2.4.1. The

original analysis employed PMI as a co-occurrence weight to measure the association between two artists A and B as:

$$\mathrm{PMI}(A, B) = \log \frac{p(A, B)}{p(A)\, p(B)} \tag{7.1}$$

acquiring a positive value when the two artists co-occur more frequently than their independent marginal probabilities would predict (Bouma 2009, Amancio 2016, Berahmand et al. 2024). However, it is known to be sensitive to low-frequency pairs, meaning that when two artists who appear together only once but rarely appear individually receive an artificially inflated score. To address this, we complemented PMI with the Dice coefficient (Manning and Schütze 1999, Jost and Jost 2019):

$$\mathrm{Dice}(A, B) = \frac{2\,|\,A \cap B\,|}{|\,A\,| + |\,B\,|} \tag{7.2}$$

which normalizes co-occurrence frequencies relative to each artist's marginal frequencies, yielding a more symmetric and frequency-robust measure of association. Combined, PMI captures statistically surprising associations, while Dice captures proportionally stable ones.

As shown in Santini et al. (2024), having access to the Index of Names, a formal list of people that can serve as GT, enables the identification of individuals already known to be in the text using an existing NER system (cf. Section 5.2.4.1). Still, PMI and the Dice coefficient only estimate relational structure; this is why the rest of the dissertation takes a different approach. Co-occurrence is mainly a proximity measure that notes when two entities appear in the same part of the text (Amancio 2016). It cannot tell what kind of relationship exists: for example, was Michelangelo a mentor to Raphael, were they rivals, or were they just mentioned together by chance? PMI does not account for mention context and cannot identify entities or relationships that are only implied. The extraction pipeline ObliquER, described in Section 7.4, is designed to address these issues by shifting the focus from statistical co-occurrence to KE.

7.2.2 Distribution across volumes

Figure 7.1 and Figure 7.2 show the page-level co-occurrence counts per volume for each measure. The distribution of co-occurrences across the ten volumes of de Vere's translation is uneven for both. We report both because each captures a different facet of artistic co-mention, and comparing them clarifies the relational structure. For PMI, the statistics show that Volume 6 contains the highest number of co-occurrences, corresponding to the biographies of the most prominent figures of the High Renaissance. Volumes 1 and 10 – covering the earliest artists and supplementary material, respectively – contain significantly fewer. The Dice coefficient shows a more refined perspective; however, the distributions look very similar. Both peak in Volume 6 (Dice: 135, PMI: 125), both drop in Volume 10 (Dice: 8, PMI: 6), and both show a dip in Volume 3. Dice

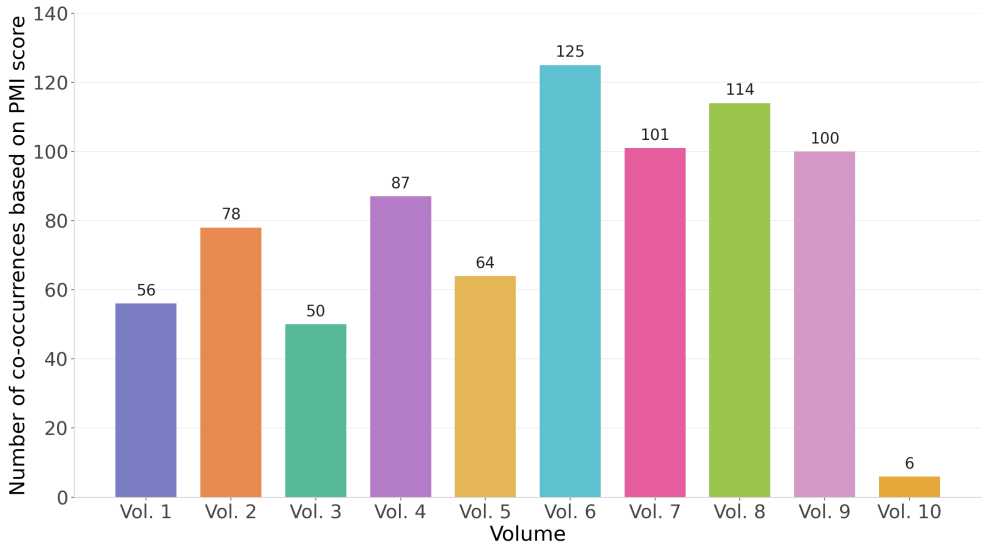


Figure 7.1: Distribution of PMI-based co-occurrences per volume in de Vere's translation of *The Lives*.

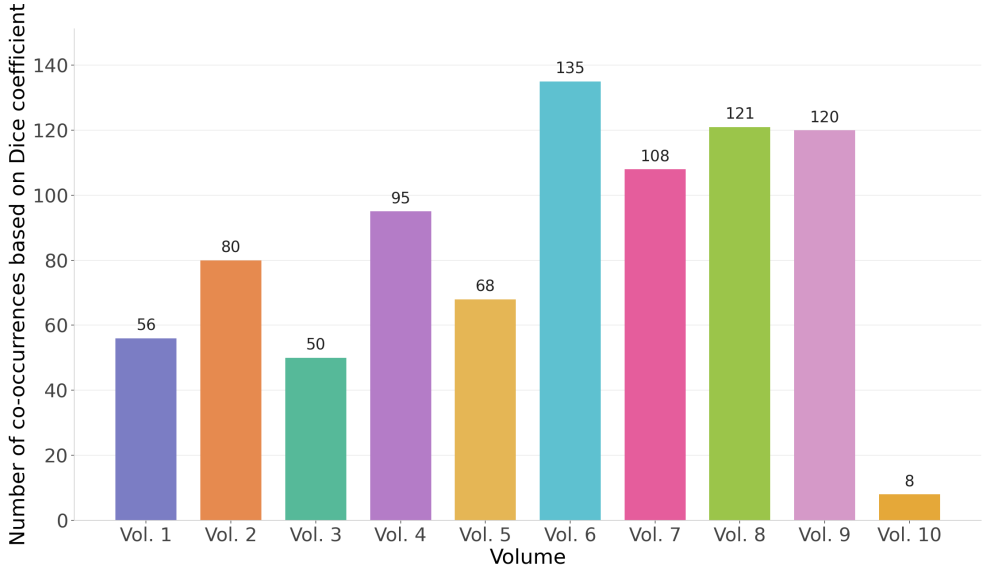


Figure 7.2: Distribution of Dice coefficient-based co-occurrences per volume in de Vere's translation of *The Lives*.

keeps slightly more pairs in every volume, with 841 events in total compared to 781 for PMI, and this difference grows in the later volumes (Vol. 9: 120 vs. 100). This matches what we expect, with PMI using logarithmic normalization, which reduces the impact of rare associations, while Dice, as a proportional overlap measure, keeps them if their relative co-occurrence is high. For the methodology, this means that evaluation across a single volume would not be representative; sampling must span multiple volumes to capture the full range of narrative density, as described in the annotation strategy of Section 7.4.4.

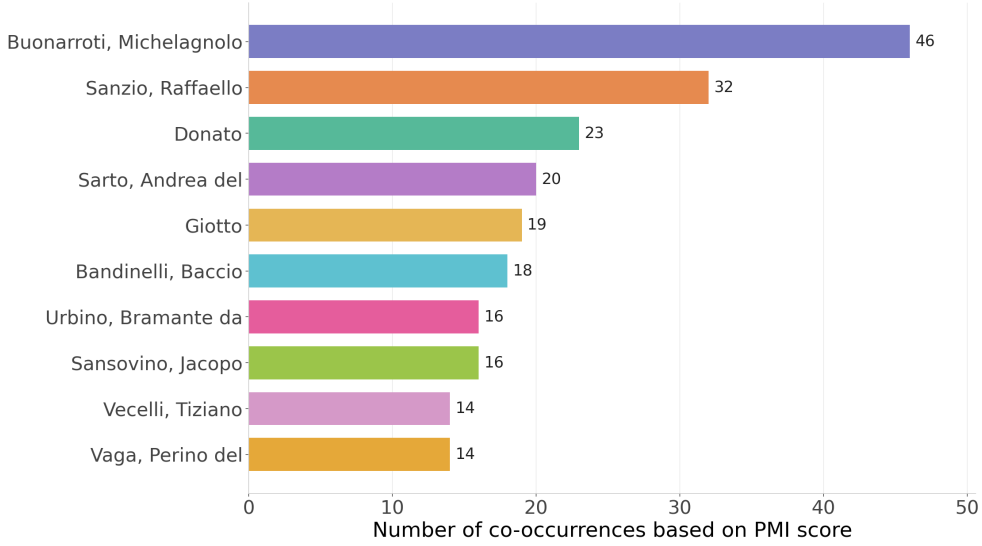


Figure 7.3: Top ten artists ranked by number of PMI-based co-occurrences across all volumes of de Vere's translation of *The Lives*.

The differences between the measures become clearer when looking at individual artists. Figure 7.4 and Figure 7.3 list the ten artists with the most co-occurrence events for each measure. Nine of the ten names are the same in both rankings: Michelangelo, Raphael, Donatello, Andrea del Sarto, Giotto, Bandinelli, Bramante, Sansovino, and Titian. Their scores differ, however. Michelangelo stands out much more with Dice (66) than with PMI (46), and a similar pattern appears for Raphael (39 vs. 32). This shows how the two measures work, since Dice favors artists who are often mentioned together, so those with high frequency and many co-mentions get higher scores. PMI reduces the effect of frequent mentions and highlights more specific associations. Both measures identify the same main group of artists. This small cluster of figures dominates and has already been recognized in art-historical literature on Vasari's work as canonical (Campbell 2008). This imbalance affects both extraction and evaluation. Pipelines that focus

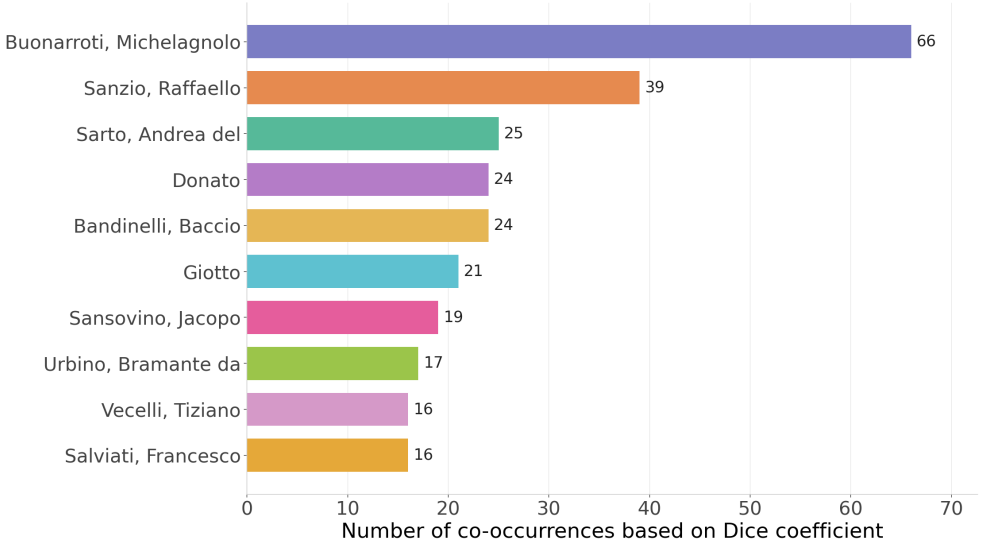


Figure 7.4: Top ten artists ranked by number of Dice coefficient-based co-occurrences across all volumes of de Vere's translation of *The Lives*.

on high-frequency entities may not perform well on the many lesser-known people, even though these make up most of the text. We discuss this distributional asymmetry further in Section 7.4.4, which describes the annotation strategy for implicit entity recognition.

This analysis is exploratory and not a main contribution of the dissertation. Santini et al. (2024) started this line of research; we have repeated it using the extracted-paragraph view, added the Dice coefficient to the original PMI analysis, and found both approaches to be consistent. We do not proceed with network analyses based on these counts because, as discussed earlier, co-occurrence only indicates proximity and cannot reveal the specific types of relationships necessary to understand the layers of Vasari's work. The co-occurrences themselves are nevertheless represented in the final KG as a dedicated co-occurrence class, the initial separation of which in the Viewsari ontology this analysis motivated (cf. Section 7.3); the PMI and Dice scores remain an exploratory layer over those pairs rather than weights materialized on the graph. The same distributional findings were supplementary to the design of the GT corpus and its sampling strategy. The canonical skew toward a small cluster of high-frequency artists directly motivated the stratified sampling of Section 7.4.4, which caps the over-represented biographies and redistributes the annotation budget toward the long tail of lesser-known figures so that the evaluation set is not itself dominated by the same few canonical names. From there, ObliquER (Section 7.4) becomes the focus for extracting further information from the text.

7.2.3 Preliminary comparison of explicit entity type coverage

In addition to the co-occurrence analysis, as mentioned in Section 5.2.4.2, prior work has produced ExtrART, an evaluation dataset for IE, compiled from the same edition of the text.

Coverage across entity types is considerably uneven, as shown in Table 7.1. Aside from the traditional named entity types, the extraction of artworks and motifs remains comparatively sparse, with 133 and 311 instances, respectively. This disparity reflects a fundamental difference in how these entity types surface in the text. Persons, for example, are predominantly named explicitly, whereas artworks and motifs are frequently introduced through descriptive or iconographic language that resists surface-form matching. A painting may be referred to as "the panel he executed for the high altar," and a motif as "the scene of the Annunciation in the upper register," neither of which corresponds to a canonical label retrievable by standard NER.

Crucially, ExtrART itself does not attempt to resolve this problem, remaining anchored in explicit textual mentions. The entity types it covers best are precisely those adhering to surface-form identification, whereas the sparser categories are those in which reference is most frequently implicit. The present work directly addresses this gap. We propose the ObliquER pipeline (Section 7.4) to extend entity recognition beyond explicit surface forms into the domain of implicit entity extraction, treating this as an epistemic act in which generative extraction tools are understood as instruments for interpreting, rather than merely retrieving, the entities latent in the text. The comparison across artwork entities will be revisited and substantially extended in Section 7.4, where implicit mentions are incorporated into the analysis.

7.2.4 Prototypical knowledge graph

The first prototype of a KG by Santini et al. (2022, 2024), introduced in Section 5.2.4.1, is sufficient for exploratory social network analysis but exposes several limitations that require addressing. All co-occurrence edges are of the same type, and co-occurrences are not anchored to the paragraphs from which they were derived, making it impossible to trace a relationship back to its textual source. These limitations motivate the ontology engineering process described in the following section, which introduces a formalization of relationships and provenance that supports both the existing co-occurrence data and the richer entity types to be extracted in subsequent pipeline stages.

Table 7.1: Coverage of entity types across the ExtrART reference dataset (corpus with 55 paragraphs, 27 chapters).

Entity Type	**Count**
Named Entities	
Persons (`PER`)	732
Locations (`LOC`)	414
Organizations	56
Miscellaneous	51
Artworks	
Artwork references	133
Iconographic Motifs	
Subjects (`SUBJ`)	281
Predicated states (`STATE`)	30
Span relations	51
Artistic Terminology (AAT)	
Architectural objects (`OBJarch`)	279
Generic objects (`OBJ`)	179
Visual art objects (`OBJvis`)	149
Materials (`MAT`)	76
Techniques (`TECH`)	57
Temporal Expressions	
Specific dates (`DATE`)	42
Time intervals (`INTERVAL`)	14
Total annotations	**2,544**

7.3 The Viewsari ontology: design process and iterations

This section outlines the methodology employed to design the Viewsari ontology and presents the actual design process. Supplementary materials are accessible in the GitHub repository.[2] The ontology is contribution **E1**, which operationalizes **C1** through the three-layer architecture, **C2** by closing the provenance gap at the representational level – reifying the extraction agents, prompt templates, and runs as first-class structure linked to every extracted assertion – and **C3** through the typed activity classes that frame this provenance as hermeneutic, recording the interpretive act of extraction rather than the model's internal computation.

Author contributions. The Viewsari ontology was developed collaboratively within the ISE group at FIZ Karlsruhe and KIT. Harald Sack and Jörg Waitelonis provided conceptual guidance and supervision throughout all design phases. Tabea Tietz and Etienne Posthumus contributed to the ontology engineering process, particularly in modularization. The work on ODPs was developed jointly with Bruno Sartini, Marieke van Erp, Andrea Poltronieri, Pasquale Lisena, and Harald Sack, and is presented in a corresponding publication (Ondraszek et al. 2024c).

Introduction to the Viewsari ontology. Given the interdisciplinary nature of the project, spanning art history, DH, and knowledge engineering, and the need for flexible, iterative adaptation as new requirements emerge from the source material, the XD methodology was identified as particularly suitable for developing the Viewsari ontology (Presutti et al. 2009, Blomqvist et al. 2016, Ondraszek et al. 2024a). XD was selected over alternative ontology engineering methodologies – notably NeOn (Suárez-Figueroa et al. 2015) and Methontology (Fernández et al. 1997), both situated within the broader survey of engineering methodologies in Section 5.2.1 – for three reasons specific to the requirements of this work. XD is built around the systematic reuse of ODPs, which aligns with the integration of multiple established vocabularies, such as FaBiO and PROV-O, required by the three-layer architecture described in Section 7.3.1.2 and Section 7.3.2.2. Furthermore, the focus on iterative, test-driven development through CQs provides a concrete evaluation protocol that aligns with the empirical validation of the Vasari case study, namely by analyzing the answerability of CQs and domain-expert feedback (Chapter 9). The emphasis on modular ontological commitments aligns with the requirements for representing partial, context-dependent annotations on textual mentions, since the Viewsari ontology aims to surface the different components of the textual source. A comparative survey of ontology

[2] GitHub repository for Viewsari: `https://github.com/ISE-FIZKarlsruhe/viewsari-kg/tree/main`.

engineering methodologies lies beyond the scope of this dissertation; readers seeking such an overview are referred to Iqbal et al. (2013).

Viewsari is positioned between the domain and application levels in the ontology typology (Chapter 3). It models art-historical KE outputs and captures the extraction and provenance needs of the ObliquER pipeline. As an application-focused ontology, it represents extracted content and related provenance statements, as shown in *The Lives* use case. The first version includes class definitions based on domain needs, such as identifying a person as an extracted entity or noting co-occurrences between artists. Later development phases add statements about how generative extraction creates entities, how LLMs act as interpretive agents, and how the process provides provenance information. The ontology was built in Protégé[3] during both design phases and is serialized in OWL/RDF. The final version is published under a CC BY 4.0 license in the GitHub repository linked above.

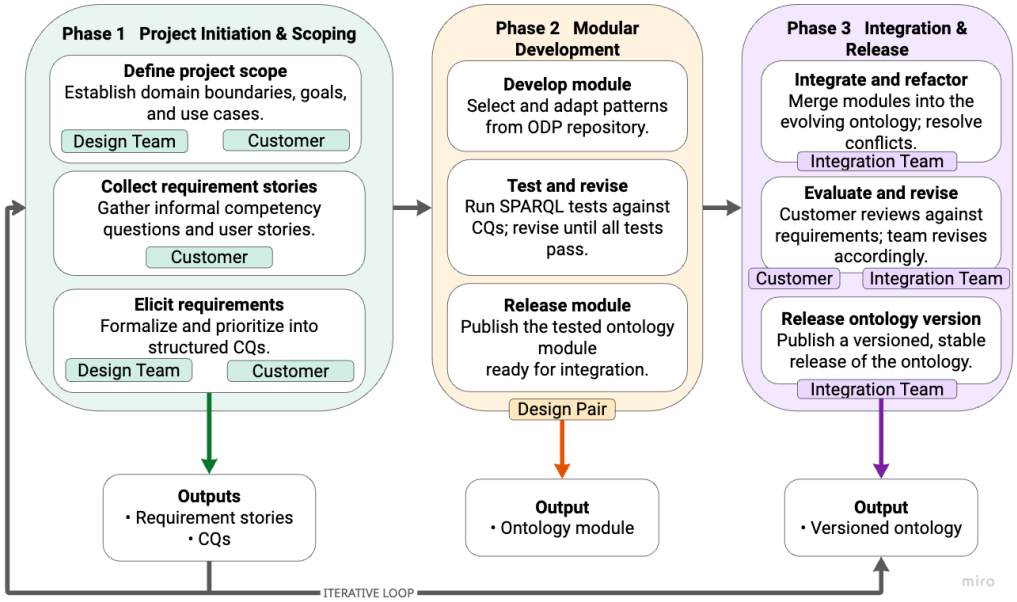


Figure 7.5: Overview of the three phases of the XD methodology, following the steps given in Blomqvist et al. (2016). The phases span project initiation and scoping, the iterative module development loop, and the integration loop. Outputs of each phase are shown below the respective phase. The iterative loop arrow indicates that newly released ontology versions feed back into the development process as new requirements emerge.

[3] Protégé: `https://protege.stanford.edu/`.

Figure 7.5 presents the three phases of XD (Blomqvist et al. 2016), starting with initiation and scoping. This phase establishes foundational agreements, including role distribution, the technical environment, representation languages, and the ODP catalog. It also produces an initial set of user stories, short, customer-driven narratives that describe specific domain tasks the ontology must support; these serve as input for all subsequent development. We refer to this phase as *requirements engineering* (cf. Section 7.3.1); we include two iterations of this phase (cf. Section 7.3.1.1 for the first run and Section 7.3.2.1 for the second). The second phase involves module development (cf. Section 7.3.2), where ontology designers iteratively select a story, elicit formal requirements as CQs, contextual statements, and reasoning requirements, and then develop and test modules by reusing and specializing patterns. SPARQL queries evaluate the CQs on representative test data. Modules are released only after domain experts confirm that all requirements are met. We refer to this as the *module development* phase and, once again, separate it into two iterations (cf. Section 7.3.1.2 for the first run and Section 7.3.2.2 for the second). The third phase focuses on integration, where the team merges released modules into the evolving ontology, resolves conflicts using reasoning techniques, and publishes the latest version. Domain-expert feedback is essential at each phase to ensure that all requirements, including the correct use of domain terminology, are met before release. Given that this methodology focuses on iterations, the integration loop feeds back into the development loop, providing refinements and new requirements to module teams, with continuous emergence of new requirements and ontology modifications as modules are integrated and evaluated.

7.3.1 Viewsari ontology design phase I

This section traces the first iteration of the Viewsari ontology, beginning with the requirements elicited from art-history domain experts and the user stories distilled from them (Section 7.3.1.1). It then maps these requirements onto reusable ODPs and formalizes them as the classes and properties of the first version of the ontology (Section 7.3.1.2).

7.3.1.1 Viewsari phase I: requirements engineering

Author contributions. Requirements elicitation for this phase was conducted jointly with the domain experts, including Prof. Dr. Anna Schreurs-Morét, Dr. Grischka Petri, Dr. Ulrike Blumenthal, Olivia Schmidt-Thomee, Paula Schulze, Annika Leuschner, and Brechje Bakker, and was coordinated by the author (cf. Section 7.3.1.1). The team subsequently supported decisions in the second phase as well (cf. Section 7.3.2).

The Viewsari ontology was motivated directly by the limited expressivity of the KG in prior work (Santini et al. 2022): only structured, semantically rich data enables advanced querying, reasoning, and interoperability.

Domain scoping. The initial domain set was apparent, with *The Lives* representing the subdomain, while art history and the Italian Renaissance are the super-domains. Because all information comes from Vasari's writings, the scope is best described as Vasari's Renaissance. The text contains biographies that share details about artists, their lives, related artworks, motifs, and locations, and it includes relational information as described in Santini et al. (2022). The first version of the ontology starts with the social network described in Santini et al. (2022, 2024), so the main focus is on people and their relationships. This information has shaped the ontology's design, especially in defining entity types, patterns, and other conceptual elements.

Expert consultations. To elicit requirements systematically, we invited a group of art history domain experts to a series of collaborative sessions. The first group comprised three art history students from Trier University, a trained art historian working in library science and media and communication, and an art historian and lawyer specializing in art history and copyright law, with extensive expertise in the Italian Renaissance, particularly around Vasari (as noted in the introductory paragraph). The group was later expanded to include Prof. Schreurs-Moret, professor of early modern art history at Freiburg University, and her two colleagues, Paula Schulze and Olivia Schmidt-Thomee, who provided expert feedback on the GT corpus and the choice of texts (cf. Section 7.4.4).

Throughout the meetings, we presented both the existing social network and the first KG, making the limitations of the current representation explicit and using them as a starting point for discussion. Requirements were elicited and organized thematically on a shared Canva and Miro board,[4] covering topics such as artists and their collaborations, artworks, locations, historical events, provenance and metadata modeling, linked data enrichment, user interaction, and visualization.

Domain experts emphasized several priorities, most notably the need for dedicated modules to represent provenance, typed person-to-person relations, and co-occurrences, as well as the contextual grounding of extracted entities. They further emphasized the importance of representing historical events and their temporal context, enabling the placement of artists, artworks, and relationships within historical developments. Additionally, the ability to cross-reference extracted knowledge with external sources, thereby supporting fact-checking and federated queries across

[4] Canva board for requirements engineering: `https://www.canva.com/design/DAF6nI8fbmY/N2fAZuySC_EjF5Cw-mZsJQ/edit`.

datasets such as Wikidata (Vrandečić and Krötzsch 2014), was identified as a key requirement to ensure the KG's extensibility.[5]

In particular, provenance emerged as a core requirement: the experts' recurring response to any representation that presented extractions as facts, such as a co-occurrence between artists, was to ask, "how are we to believe this?" This kind of feedback surfaced the importance of source criticism and framed traceability and context sensitivity as essential for interpreting results from entity-extraction frameworks. Because all extracted resources and their relationships come from one or more paragraphs in *The Lives*, the KG must record the provenance of each piece of information (Galka and Vogeler 2025, Yang et al. 2026). Table 7.2 summarizes our requirements.

During the meetings, requirements were collected in the form of general research scenarios, user stories, and ad hoc, supplemented with personas. Personas group related stories into a non-structured text of up to 250 characters on a Canva board.[6] Each persona has one research scenario and multiple user stories. In total, we constructed four personas representing a traditional researcher in art history (15 user stories), a student (6 user stories), a professor in digital art history (4 user stories), and a computer scientist (7 user stories). An example user story from the art historian persona is the following:

> As a researcher, I want to explore co-occurrences of artists mentioned in *The Lives* so that I can identify potential collaborations.

From these stories, we formalized 39 CQs during the first phase of requirements engineering. Ten of these are functional questions covering the Viewsari web application – search, visualization, and exploration – that concern the system layer and are discussed separately in Section 7.7 rather than being entered into the ontology catalog. The remaining 29 constitute the Phase I catalog (Section A, `CQI.1-29`); as the catalog's *Type* column records, 25 are SPARQL-answerable (`select`/`ask`) and four are descriptive questions concerning data access and reuse. To exemplify this procedure, the user story introduced above yields the following CQs: *Who co-occurred with whom? Did artist XY co-occur with artist YX? Where in the text did artists XY and YX*

[5] The choice of Wikidata over the Getty vocabularies for artworks is motivated by coverage since Vasari's *The Lives* mentions artworks across a wide spectrum of canonicity, ranging from canonical masterpieces to lost works. Getty, while authoritative, has comparatively limited coverage of individual artworks, whereas Wikidata's crowd-sourced, cross-domain graph offers broader (if uneven) representation of artworks, patrons, locations, and biographical entities. Additionally, Wikidata maintains explicit external-identifier properties for Getty, so a Wikidata-anchored linking does not preclude later alignment with Getty resources; rather, it positions Viewsari to bridge both ecosystems (Zhang et al. 2022, Centelles Velilla and Ferran-Ferrer 2025). The methodological trade-off is that Wikidata entries vary in completeness and authoritativeness compared to curated vocabularies; this is partially mitigated by the provenance modeling described in Section 2.4, which records the linking activity and the software agent responsible for each extraction, thereby leaving the link transparent.

[6] Canva board for requirements engineering with the domain experts: `https://www.canva.com/design/DAF6nI8fbmY/N2fAZuySC_EjF5Cw-mZsJQ/edit`.

Table 7.2: Requirements for Viewsari collected during the initial requirements elicitation phase.

Requirements
Data and provenance
Representation of persons (artists) and all other extracted entities types – including organizations, places, artworks, motifs, terms, and dates – as structured resources.
Typed relations between all entity types, including persons, artworks, and locations (e.g. mentorship, co-occurrence).
Provenance and positional information for all extracted entities and their relations, connected to the source paragraph, biography, and top-level bibliographic metadata of the work.
Direct links to original text passages from *The Lives*, including the Italian source text, and to external authority data such as Wikidata.
Relevant attributes of resources: birth and death dates for persons, creation dates for artworks.
Representation of historical events and their temporal context, connected to relevant persons, artworks, and locations.
Support for fact-checking and federated queries across external datasets such as Wikidata.
User-facing functionality
Index searches for extracted resources: artists, artworks, locations, etc.
Semantic and visual searches
Exploration of resources and co-occurrences without a pre-formed RQ
Visibility of contextual implications and the connection strength between resources
Support for interpretation and discovery through the visualization of resources and their connectedness

co-occur? From the CQs, a foundational ontology for Viewsari was constructed. The CQs serve as the baseline for ontology evaluation in subsequent steps of the project; the parallel functional requirements summarized in Table 7.2 instead feed the system layer.

7.3.1.2 Viewsari phase I.2: module development

With the requirements established, the next step was to identify existing ODPs that would serve as building blocks for the modules. The requirements table (Table 7.2) was grouped into three distinct representational clusters, each corresponding to a module in the first version of the Viewsari ontology, namely a *bibliographic layer* capturing the metadata of *The Lives* (e.g., versions, editions, translations, online copies, ...), a *structural layer* representing the document hierarchy and the paragraph-level anchoring of extracted content to link back provenance, and a *content layer*, modeling exactly that extracted content, including relationships. This three-module architecture has remained the stable backbone in the subsequent iteration.

It should be noted that the ontology was still in its exploratory phase at this point, and we did not use a top-level ontology to anchor the classes and properties. This was added in the second phase.

Reused and adapted ontology design patterns. The content layer reused four content patterns from the ODP repository,[7] each addressing one recurring modeling problem from the requirements. The *Persons* content pattern[8] was instantiated as `viewsari:person` for artists and other named individuals from the Index of Names; the pattern was taken as a direct reference for the class implementation and required no further specialization at the pattern level. This pattern was relevant given its definition of persons as a "commonsense intuition, which does not apparently distinguish between either natural or social persons." Since we had persons in the dataset that could not be fully linked to existing KBs, this helped reduce the constraints around someone having to be 'real' to be considered a person. In parallel, the *Place* content pattern[9] was reused with its class `:Place` renamed to `viewsari:location` to align with the dominant terminology in the requirements catalog, while its relational shape – `:hasLocation`, `:isLocationOf` – was retained unchanged. This was used predominantly to situate extracted content within specific locations mentioned in Vasari's work.

[7] ODPs repository: `https://odpa.github.io/patterns-repository/`.

[8] Person content pattern for classes: `https://web.archive.org/web/20250114034150/http://ontologydesignpatterns.org/wiki/Submissions:Persons/Person`.

[9] Place content pattern: `https://web.archive.org/web/20250211050648/http://ontologydesignpatterns.org/wiki/Submissions:Place`.

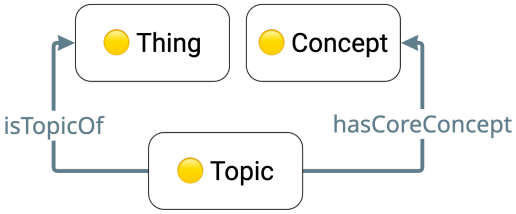


Figure 7.6: The Topic ODP, adapted from the ODP repository. A `Topic` mediates between a `Thing` (the resource being indexed) and a `Concept` (the semantic category serving as its subject), related through the inverse properties `isTopicOf` and `hasCore` (partially visible at the right edge of the figure). The pattern factors topical aboutness out of direct binary links, allowing additional metadata to be attached to the indexing relation itself.

Furthermore, the *Topic* content pattern[10] was instantiated for motifs, genres, and iconographic subjects available in ExtrART. As Figure 7.6 shows, the pattern can be used to add topics to undefined things, and topics can be specialized using concepts.

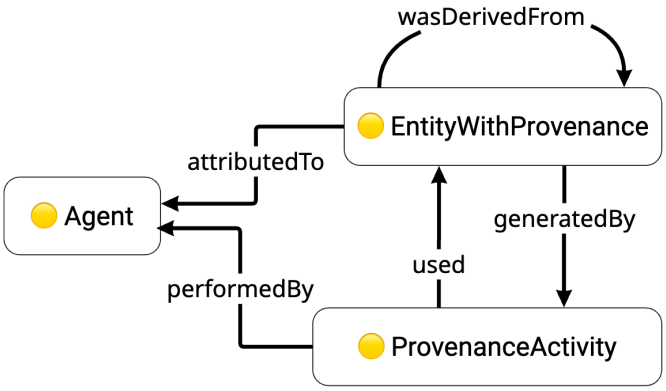


Figure 7.7: Top-level classes of the Provenance ODP, adapted from the ODP repository and Shimizu et al. (2019): an `EntityWithProvenance` for which provenance is to be recorded, an `Agent` responsible for the assertion or generation, and a `ProvenanceActivity` representing the process through which the entity came about. The triad abstracts the core structure shared by PROV-O and related provenance vocabularies (Lebo et al. 2013).

The *Provenance* content pattern[11] provided critical inspiration for representing source evidence for extracted co-occurrences (Shimizu et al. 2019). Figure 7.7 shows the central idea that an entity is derived from some other entity, generated by some activity, and associated with some agent. We adapted it to the co-occurrence-in-paragraph context, which is extended further in the latest ontology version (see Section 7.3.2.2). In this version, with classes shown in Figure 7.9,

[10] Topic content pattern: `https://web.archive.org/web/20241202223449/http://ontologydesignpatterns.org/wiki/Submissions:Topic`.

[11] Provenance pattern: `https://web.archive.org/web/20250316121248/http://ontologydesignpatterns.org/wiki/Submissions:Provenance`.

the provenance pattern described the reification of artist co-occurrences as first-class resources, whose grounding was made explicit through the `:in` relation to a `:Paragraph` and, transitively via `:on` and `:in`, to the containing `:Page` and `:Volume`. The participating artists were attached via `:involved resource`, while the textual chain anchored each assertion to a specific locus in the physical-textual hierarchy of *The Lives*. The generating `ProvenanceActivity` and responsible `Agent` were not yet modeled at this stage but are introduced in the refinement discussed in Section 7.3.2.2.

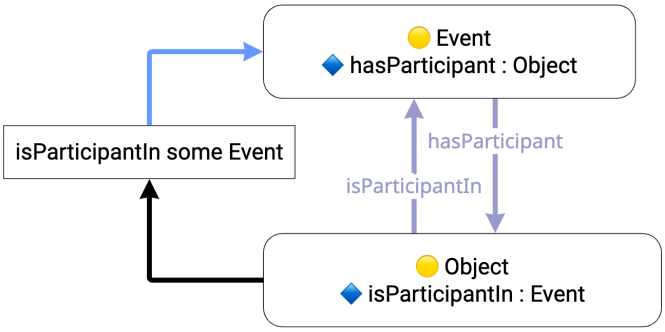


Figure 7.8: The Participation ODP, adapted from the ODPs repository. An `Event` and an `Object` are related through the pair of inverse properties `hasParticipant` and `isParticipantIn`, with the class restriction `isParticipantIn some Event` expressing that any object qualifying as a participant must take part in at least one event.

A fifth, structural pattern, the *Participation* pattern,[12] shown in Figure 7.8, provided the structural template for the first representation of co-occurrences as first-class entities. The rationale, the earlier reification and singleton-property attempts that preceded this choice, and the resulting `viewsari:cooccurrence` class are described below under the content-layer classes. We show the derivation in Figure 7.14.

With the pattern catalog established, the next step is to formalize the modules as classes and properties. Each key modeling decision is grounded in an explicit CQ; the co-occurrence class, for instance, addresses CQ3: "Which artists are mentioned together in the same passage?" A detailed Graffoo[13] visualization of the classes and properties is available in the GitHub repository.[14]

[12] Participation pattern: `https://web.archive.org/web/20250219135335/http://ontologydesignpatterns.org/wiki/Submissions:Participation`.
[13] Graffoo: `https://essepuntato.it/graffoo/`.
[14] Formalization of the first ontology version: `https://github.com/ISE-FIZKarlsruhe/viewsari-kg/tree/main?tab=readme-ov-file#formalization`.

Resulting classes. The classes are organized along the three layers introduced above, beginning with the bibliographic layer. An overview of this is shown in Figure 7.9, with the first version of the ontology developed across these three axes.

Bibliographic layer. The variance across editions of a written work was represented in the first Viewsari ontology via the bibliographic layer, which focused on classes representing literary works and their embodiments. One of the central classes in the bibliographic layer was the `viewsari:book` class, which was designed to represent the conceptual, abstract idea of *The Lives*, standing above specific real-world versions of it, and thus, harmonizing the idea of the work in exactly such a scenario. In parallel with the FRBR concept of a 'Work', it signified the formal representation of abstract intellectual content, independent of any particular form. Introducing this class into the Viewsari ontology met the requirements, as it provides a formal representation of Vasari's original text while distinguishing it from its various editions and translations.

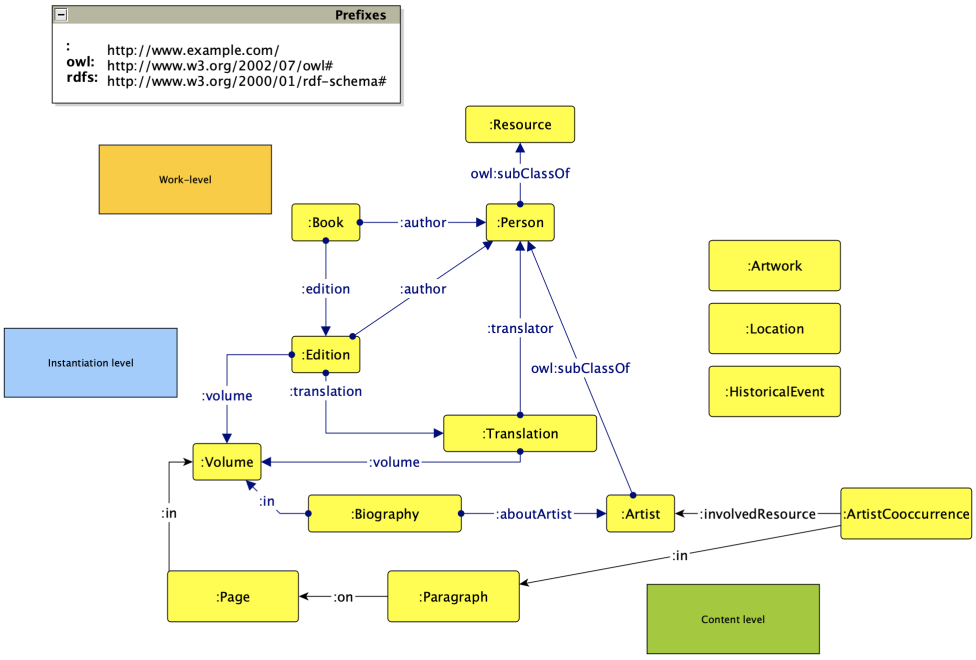


Figure 7.9: Schematic overview of the Viewsari ontology. The three layers–Work, Instantiation, and Content–separate bibliographic identity (`:Book`/`:Edition`/`:Translation`) from physical-textual location (`:Volume`/`:Page`/`:Paragraph`) and from interpretive content (`:Biography`, `:Artist`, `:ArtistCooccurrence`). Artist co-occurrences are reified as first-class resources anchored to specific paragraphs, enabling provenance back to the textual locus of each relation.

The bibliographic layer also included classes representing instantiations of *The Lives*, specifically `viewsari:edition`, `viewsari:translation`, and `viewsari:volume`, enabling the tracing of editorial provenance. `viewsari:edition` denoted a published version, such as the 1568 Italian edition. `viewsari:translation` captured translations into other languages; for example, Gaston du C. de Vere's 1912 English translation, which serves as the primary source for the current Viewsari KG, was an instance of this class. `viewsari:volume` represented physical divisions within an edition or translation; de Vere's translation is divided into ten volumes, each containing subsets of artist biographies. These classes carried metadata via properties such as `viewsari:author` or `viewsari:translator` and linked to the work level through properties like `viewsari:edition of` and `viewsari:translation of`.

This structure lets multiple versions of the text co-exist and links extracted content back to specific passages and editions of *The Lives*, enabling comparison of annotations across editions for textual analysis and fact-checking against multiple sources. For instance, the following SPARQL query demonstrates how this model enables comparison of artist co-occurrences across different translations. It retrieves all paragraphs in which two specific artists are mentioned together, listing the edition or translation in which the co-occurrence appears. This allows a researcher to immediately see whether a given association between artists is unique to a particular edition or consistently represented across translations, highlighting potential discrepancies or interpretive differences.

```
SELECT ?edition ?volume ?paragraph ?artist1 ?artist2 WHERE {
 ?cooccurrence a viewsari:Cooccurrence ;
        viewsari:mentions ?artist1, ?artist2 ;
        viewsari:mentionedIn ?paragraph .
 ?paragraph viewsari:inVolume ?volume ;
       viewsari:inEdition ?edition .
 FILTER (?artist1 = <ARTIST1_URI> && ?artist2 = <ARTIST2_URI>)
}
```

This illustrates how the structure supports analytical questions about the textual transmission of relationships, and offers a concrete mechanism for examining historical narratives across different versions of *The Lives*.

Structural layer. To further address positional provenance requirements, the ontology incorporated document-structural classes, such as `viewsari:paragraph`, which served as positional anchors for co-occurrences and entity mentions. In the typical process of scholarly transcription or citation in the humanities, a researcher might write, "As Vasari notes on page 143..." or "In

the third paragraph of the Life of Michelangelo...," making exact positional references crucial for traceability. Each paragraph was identified by its position within the document hierarchy (volume, biography, page) and contained the text snippet from which entities were extracted. Furthermore, `viewsari:page` represented individual pages within a volume, providing an intermediate level of document structure. The `viewsari:biography` represented the biography of a single artist.

The modeling of provenance information in the first version of the Viewsari ontology was inspired by existing ontologies and vocabularies, including PROV-O (Lebo et al. 2013). In the initial version presented here, direct mappings to these frameworks were not yet formalized but are available in the current version, which we will explain in Section 7.3.2.2.

Content layer. The super-class `viewsari:resource` is the predominant class of the content layer and was introduced to provide a common denominator for all extracted entities. This abstraction supports generic queries across different entity types. Specific subclasses were derived directly from the data requirements and CQs. A `viewsari:person` represents any person mentioned in *The Lives*. Artists, among the most important figures, receive their own class, `viewsari:artist`. The distinction between `Person` and `Artist` allows for more precise querying, for example, distinguishing between co-occurrences involving two artists versus those involving an artist and the Pope. Additionally, we add further classes for extracted entities, including `viewsari:artwork` to represent artworks mentioned in the biographies, `viewsari:location` for geographical places mentioned in the text, and `viewsari:historical event` for events described in or inferred from the text.

The co-occurrence class: A relation as a first-class entity. The most distinctive feature of the first ontology is the representation of co-occurrences as first-class entities, using `viewsari:cooccurrence`. As described in Section 7.3.1, each co-occurrence statement needs to include provenance, association strength (such as PMI or the Dice coefficient), and a paragraph anchor. This means it cannot just be an annotated edge between people; instead, it should be seen as a resource that can be studied on its own. This approach is also important methodologically. A co-occurrence is something we can directly observe through a specific extraction process, rather than a typed relation inferred by relation extraction. By making co-occurrences explicit, we keep a clear line between what is measured in the text and what is inferred about the world. The downside is that this approach is less expressive than using a typed predicate, since it only shows that two things are mentioned together, not their relationship. Two earlier designs were tried and abandoned before settling on a dedicated class. The motivation also aligns with the provenance principle developed in Section 2.4. We differentiate an observable unit yielded by an explicit

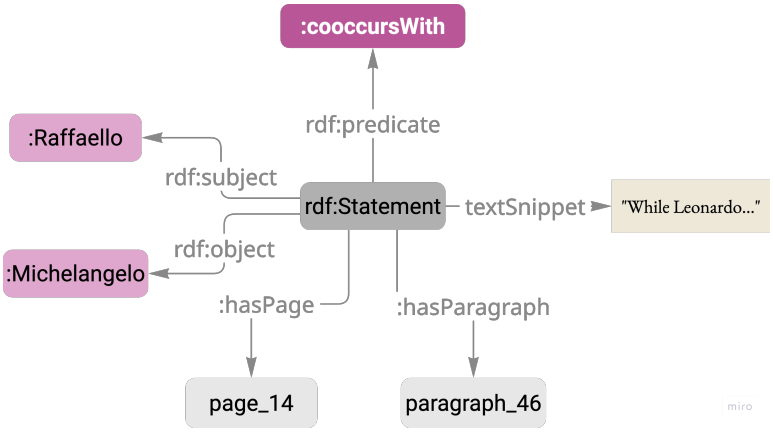


Figure 7.10: Reification of a co-occurrence assertion via `rdf:Statement`. The triple (`:Raffaelo, :cooccursWith, :Michelangelo`) is reified as an `rdf:Statement` resource carrying its subject, predicate, and object as explicit properties, which permits attaching contextual metadata–here the textual locus (`:hasPage`, `:hasParagraph`) and the supporting evidence snippet (`textSnippet`)–without committing to the asserted triple itself.

extraction procedure from a typed relation between two artists, which would be an interpretive inference that merges measurement with claim. Reifying co-occurrence as a first-class entity keeps these epistemic levels separate, at the cost of semantic expressivity relative to a typed predicate.

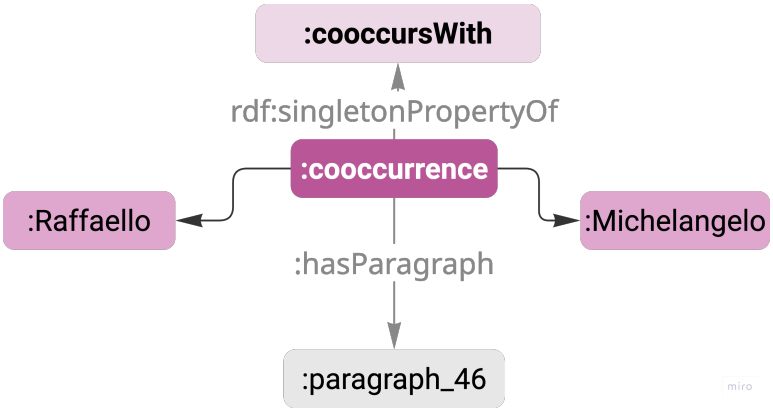


Figure 7.11: Reification via the singleton property pattern. The instance-specific predicate `:cooccurrence` is declared an `rdf:singletonPropertyOf` the generic property `:cooccursWith` and used as the predicate of the triple (`:Raffaelo, :cooccurrence, :Michelangelo`), allowing contextual metadata such as `:hasParagraph` to be attached directly to the singleton predicate rather than to a separate reifying resource.

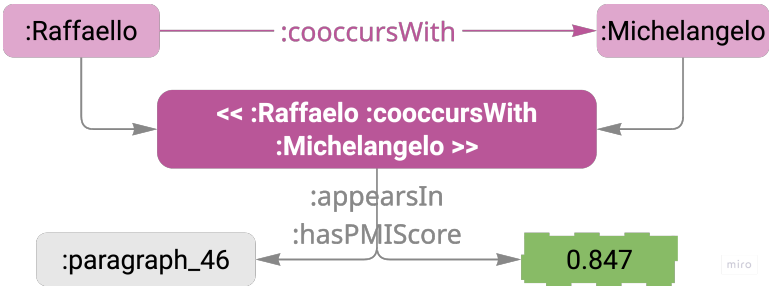


Figure 7.12: Reification via RDF-star. The asserted triple `:Raffaello :cooccursWith :Michelangelo` is itself referenced as a quoted triple (« ... ») that serves as the subject of further statements attaching its textual locus (`:appearsIn`) and a PMI score (`:hasPMIScore`), avoiding the structural overhead of classical reification while preserving access to the original triple.

The first modeled a co-occurrence between two persons as a reified `rdf:Statement`,[15] with properties such as `viewsari:text snippet`, `viewsari:has page`, and `viewsari:has paragraph` attached directly to the statement node (Figure 7.10). An intermediate iteration explored singleton properties (Figure 7.11) and a pattern inspired by RDF-star (Figure 7.12), in which a co-occurrence instance carried `rdf:singletonPropertyOf` a `viewsari:stated by` link to an algorithm instance, and a `viewsari:appears on` link to a paragraph. Both approaches require multiple joins and nested graph patterns for what should be simple queries, and neither adequately captures the semantic richness of a co-occurrence. Retrieving all artist co-occurrences with their paragraph and text snippet under reification needs several joins, whereas a dedicated class makes the same query direct and readable.

A first attempt promoted these edges to typed binary relations (Figure 7.13), substituting predicates such as `:collaboratedWith` for the generic co-occurrence link; this preserved the lightness of an edge but discarded the provenance and association-strength information that the requirements demanded, and it forced a relation-type commitment that the textual evidence often did not support.

The preliminary solution, guided by the Participation pattern and mirroring its use in the *Musical Meetups* ontology module (Polifonia Project 2023), was to model a co-occurrence as an event with participation links to the involved persons (Figure 7.14), the participation instance carrying provenance and the association score. Treating a co-occurrence as a subclass of `Event` is licensed by the pattern's domain constraint on `hasParticipant`, given that any resource bearing participants is, by definition, an event in the pattern's sense: a perdurant entity that unfolds in time and aggregates participating objects. A textual co-occurrence fits this characterization if one reads the joint mention of two artists in a paragraph as a discourse event situated in the narrative

[15] Formalization of RDF Reification: `https://www.w3.org/wiki/RdfReification`.

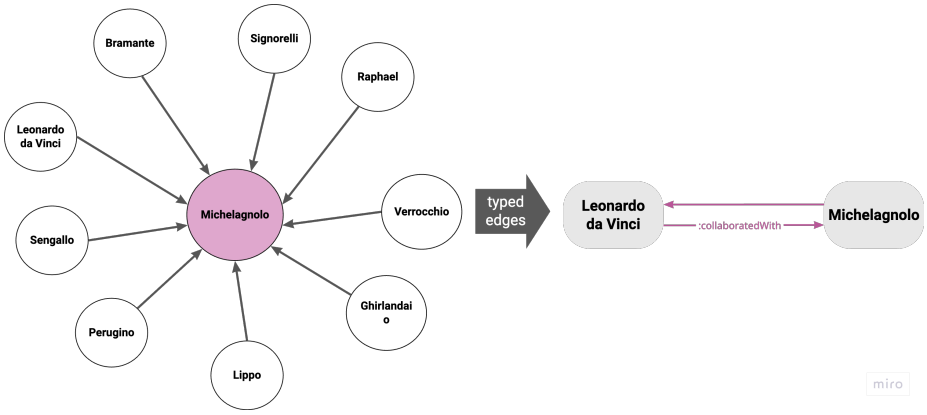


Figure 7.13: From untyped co-occurrence to typed relation. On the left, Michelagnolo is connected to other artists through undifferentiated co-occurrence edges that record textual adjacency without specifying its nature. On the right, a single pair is refined into a typed relation (`:collaboratedWith`).

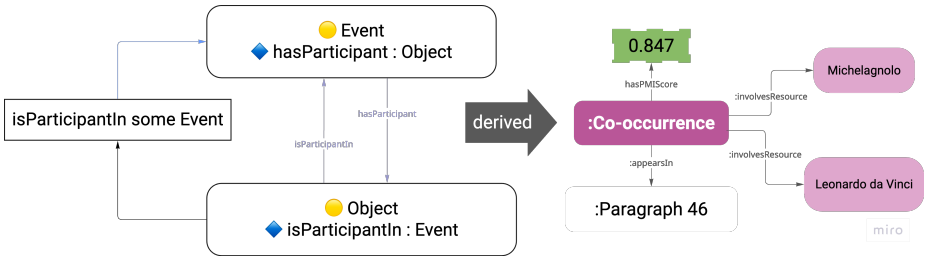


Figure 7.14: Derivation of the co-occurrence pattern from the generic participation ODP. On the left, the participation pattern relates an `Event` and an `Object` through the inverse properties `hasParticipant` and `isParticipantIn`. On the right, the Viewsari specialization reifies the joint participation of two artists (Michelagnolo, Leonardo da Vinci) in a textual unit as a `:Co-occurrence` instance, linked to its participants via `:involvesResource`, anchored to its textual locus via `:appearsIn`, and quantified by a `hasPMIScore` value.

time of the source text, with the artists as its participants and the paragraph as its spatio-textual location. We identify this constraint as metaphorical rather than ontologically committal – no claim is made that Raffaello and Michelangelo were physically co-present – but it suffices to inherit the pattern. Accordingly, because this does not align with our later understanding of the difference between textual mention and real-world referent, the Participation pattern was subsequently deprecated in the Viewsari design. However, the current ontology retains the first-class-node commitment that motivated its original reuse, on a substantially revised conceptual basis, as described in Section 7.3.2.2. These two figures show the full evolution from unlabeled social-network edges, through typed binary relations and the Participation-mediated reified form, to the present `viewsari:cooccurrence` class.

Resulting properties. Properties in the Viewsari ontology focus on intra- and inter-level object relationships, creating the connections between the three layers.

Bibliographic layer properties. At the work level, persons are assigned to a `viewsari:book` via the `viewsari:author` property, establishing authorship relationships.

```
viewsari:LeVite viewsari:author viewsari:GiorgioVasari .
```

From the work level to the **instantiation level**, the following properties establish relationships between works and their physical or digital embodiments: `viewsari:edition` connects a work to a specific edition,`viewsari:translation` connects a work or edition to a translated version, `viewsari:volume` connects an edition or translation to its constituent volumes.

Each edition or translation can carry its own authorship or translator metadata in the form of `viewsari:author`, which can be reused at this level to record editors or contributors to a specific edition and `viewsari:translator`, which records the translator of a translated edition (e.g., Gaston du C. de Vere for the English translation).

Inverse properties were added to support bidirectional navigation`viewsari:edition of` (inverse of `viewsari:edition`),`viewsari:translation of` (inverse of `viewsari:translation`, and `viewsari:volume of` (inverse of `viewsari:volume`).

These inverse properties improve queries. For example, to find all editions of *The Lives*, one can query either `?book viewsari:edition ?edition` or `?edition viewsari:edition of viewsari:le vite`, depending on the query context.

Content layer properties. At the **content level**, properties connect co-occurrences to resources and paragraphs, and connect document structures to each other: `viewsari:involved resource` (minimum cardinality 2) to link co-occurrences to the resources (typically persons or artists) that participate in it. For example, a co-occurrence between Giotto and Cimabue would have two `involved resource` links. The minimum cardinality constraint reflects the definition of a co-occurrence as requiring at least two entities. Additionally, `viewsari:in` anchors a co-occurrence to the paragraph in which it appears, providing the critical provenance link. In parallel, `viewsari:on` connects paragraphs to pages and `viewsari:about artist` connects a biography to the artist it describes, enabling queries such as “retrieve all co-occurrences mentioned in Michelangelo’s biography”.

These properties create connections between and inside of modules. For example, a query to retrieve all artworks co-occurring with a specific artist in a particular volume would traverse

several stations, beginning with the artist on the content level, to the co-occurrences via `involved resource`, to paragraphs via `in`, to pages via `on`, to volumes via `in`, all over to the desired volume.

7.3.2 Viewsari ontology design phase II

This section summarizes the second iteration of requirements engineering and ontology development for the Viewsari ontology, in which generative entity extraction with LLMs surfaces a new layer of requirements – the need to represent implicit mentions, multiple surface forms of canonical entities, and the activities that produced them (Section 7.3.2.1). It then re-models the modules accordingly, refining the bibliographic, structural, and content layers and introducing typed mention and activity hierarchies that anchor every extracted claim to its source (Section 7.3.2.2).

7.3.2.1 Viewsari phase II: extended requirements through iterative elicitation

The requirements collected during the initial user interviews, summarized in Table 7.2, constituted the first iteration of the project and focused primarily on representing persons and their co-occurrences derived from existing social network data. As the project advanced and the scope of LLM-guided generative extraction expanded, subsequent sessions with domain experts revealed additional complexities. The ontology needed not only to track the provenance of co-occurrence but also to encompass artworks, different types of mentions, and other entities. A pivotal realization during this phase was that LLMs, when used for entity extraction, do not merely retrieve explicitly stated facts from the source text but also function as interpretive agents that recognize and link implicit entities and relationships. This represents a fundamental departure from the prior work of Santini et al. (2022, 2024), in which the dataset used for the first ontology design and KG was limited to explicit entity mentions, resulting in correspondingly limited coverage of recognized entities. In contrast, the corpus analysis introduced in Section 7.2 already indicated gaps that would pose challenges for entity extraction, particularly for highly prominent entities.

Generative extraction with LLMs redefines factuality in source texts by enabling the identification of entities referenced implicitly rather than explicitly. Through collaboration with domain experts, we established that formally representing this interpretive layer is essential for the second phase of development. This methodological shift marks a broader transformation in the construction of KGs from historical and literary sources. It requires abandoning the PIEA (C1) and adopting models that encode the interpretive process as a hermeneutic provenance trail, rather than a mechanistic one (C3). As a result, the ontology must formally represent not only individual entities but also their multiple surface forms in the text, the originating text snippets, prompts, model

versions, and associated metadata. This approach addresses the provenance gap introduced by generative extraction (C2). The ontology must distinguish between implicit and explicit entity mentions, anchoring each to the source passage from which it was derived. This requirement, anticipated during the initial elicitation round, underscores the importance of positional information. Collectively, the requirements identified in this phase operationalize contributions C1 through C3 introduced earlier.

We constructed a GT that systematically distinguishes implicit from explicit mentions in *The Lives*, thereby enabling the development of prompting strategies to instruct models to extract both categories with precision. The Viewsari ontology (E1) serves as both a conceptual schema for structuring extracted knowledge and an operational guide for the extraction process, specifying the entity and relation types to be identified by the LLM. The corresponding extraction pipeline (E2, ObliquER), including the prompting methodology, annotation protocol, and approach to uncertainty, is detailed in Section 7.4.

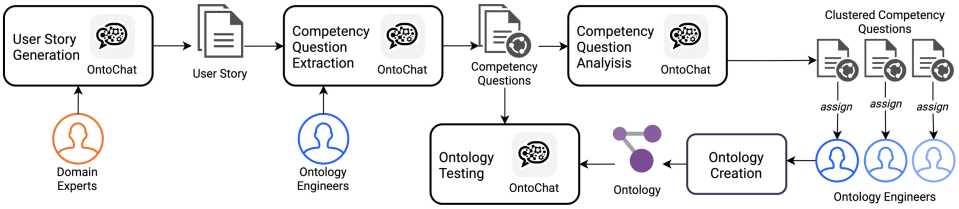


Figure 7.15: The OntoChat workflow (Zhang et al. 2024): an LLM-mediated pipeline from user-story elicitation through CQ extraction, analysis, and clustering, to ontology creation and testing, with domain experts and ontology engineers entering at distinct stages.

We reuse the four personas from the first iteration to guide the requirements engineering process. We extend the user stories with additional stories to capture the newly identified prerequisites needed to make the ontology as representative as possible. This time, we use OntoChat, a framework for conversational ontology engineering with language models (Zhang et al. 2024). As Figure 7.15 shows, there, domain experts can chat and create user stories on the go, based on requirements formulated in prompts. To this end, it is possible to set up a domain and create personas, user goals, keywords, and the like using templates, as Figure 7.16 illustrates for the persona scenarios. This also enables semi-automated CQ generation, as domain expert-designed user stories can be used directly to formulate CQs in OntoChat.

Having OntoChat as a companion for the domain experts was especially useful, as we had already completed the first phase of requirements engineering, resulting in a set of domain and exemplary

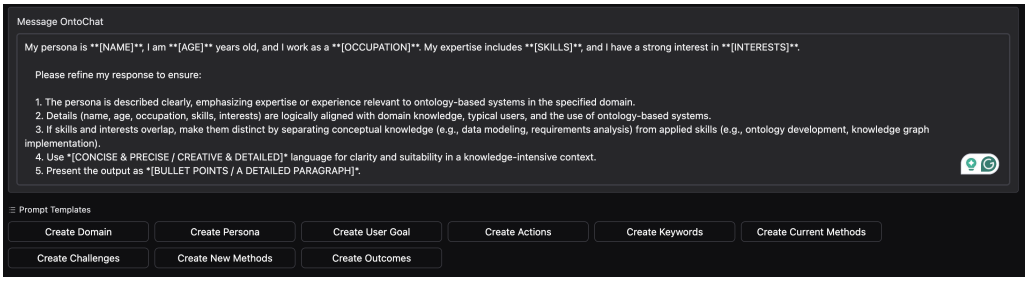


Figure 7.16: Persona-elicitation step of the OntoChat interface. A parameterized template guides the user in specifying a persona's name, age, occupation, skills, and interests, with explicit refinement criteria to ensure clarity and domain relevance. Buttons below grant access to templates for the remaining user-story components. The Markdown emphasis syntax visible in the slot placeholders (`**[NAME]**` and so on) reflects the fact that the domain experts authoring these templates are typically not experts in prompt engineering, surfacing a recurring tension in LLM-based ontology engineering between the conceptual literacy of the domain and the prompt-craft literacy presupposed by the tooling.

user stories that the domain experts can use as inspiration. The domain experts added user stories to all four personas, for example:

> As a researcher, I want to distinguish between work-level, expression-level, and manifestation-level bibliographic metadata so that I can properly cite and attribute sources across different editions and translations.

or

> As a computer scientist, I want to distinguish between different types of artist associations (statistical co-occurrence vs. explicit relationships) so that I can build accurate computational models.

In total, we collected a surplus of 15 user stories (distributed across the four personas). OntoChat generated 74 candidate CQs from them, which we consolidated into 41 distinct catalog entries (Section A, `CQII.1-41`); the higher raw count reflects finer-grained sub-questions that share a parent entry. Of these 41 entries, 27 are SPARQL-answerable (`select`/`ask`) and 14 are descriptive, as recorded in the catalog's *Type* column. Together with the 29 Phase I entries, the catalog comprises 70 CQs in total – the set against which the ontology is evaluated in Chapter 9. As in the first phase, the two user stories quoted above translate directly into catalog entries. The bibliographic story yields `CQII.19` (*At which bibliographic level (work, expression, or manifestation) is each entity of a given source recorded...?*) in the evidence-and-bibliography cluster, and the artist-association story yields `CQII.2` (*Which association types between artists – statistical co-occurrence versus explicitly typed relationship – are represented in the knowledge graph...?*) in the artists-and-relationships cluster. Not every story resolves to a graph query. Function-oriented stories yield *descriptive* entries such as `CQII.30` (*How can I use Viewsari as an interactive*

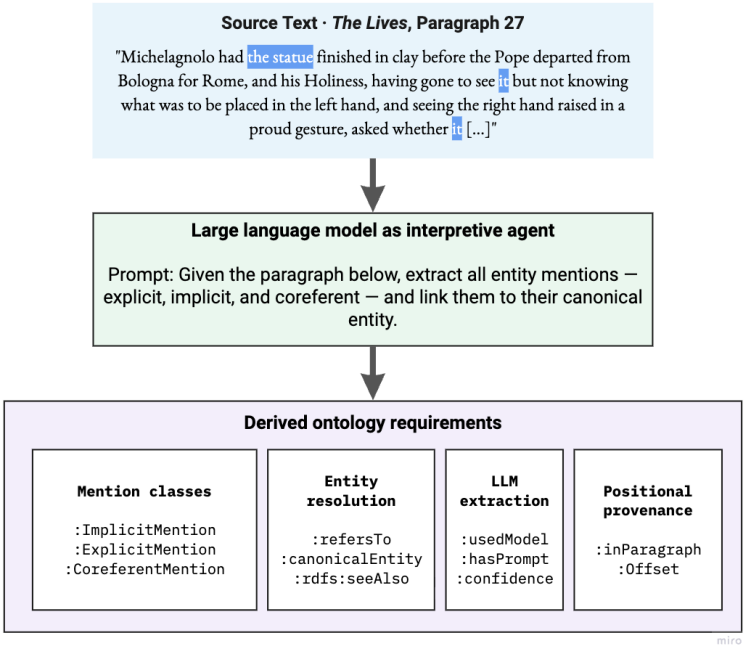


Figure 7.17: Worked example of how one paragraph of *The Lives* surfaces the requirement clusters that the Viewsari ontology has to satisfy, given the elicitation phase with domain experts.

visualization platform in my lectures...?), which the catalog answers through documentation and the web interface (Section 7.7) rather than a SPARQL pattern. These developments did not invalidate the initial elicitation but extended it across several new dimensions, each of which directly impacted ontological modeling decisions. The extended requirements are organized into thematic clusters; Figure 7.17 illustrates how the requirements pertaining to mention typology, entity resolution, and extraction provenance surface from a single paragraph of Michelangelo's biography.

1. **Co-occurrences.** Beyond co-occurrences, the domain experts emphasized the need to represent typed relationships alongside quantitative indicators such as PMI scores. Additionally, artist metadata, including names, birth and death dates, and occupations, was identified as essential, at a minimum available via federation to external KBs.

2. **Artworks.** The modeling of artwork entities constitutes a distinct requirement, necessitating the representation of attributes such as titles, creation dates, and genres. It is essential to enable systematic exploration of artworks in relation to both their creators and their textual

depictions in Vasari, particularly by federating with external KBs to facilitate interoperability. Identifying relationships among artwork instances – whether co-referring or distinct – based on the type of textual mention is critical and presupposes that the ontology can distinguish how an artwork is referred to in the text, not only that it is referred to.

3. **Evidence and bibliographic information.** Domain experts strongly reiterated the need to distinguish between work-level, expression-level, and manifestation-level metadata, and to represent dependencies between content and its source. This is particularly relevant, as we use an English translation rather than the original Italian edition.

4. **Linked data enrichment.** Providing additional information about entities from *The Lives* via Wikidata and similar authority sources shall enable fact-checking against external knowledge. SPARQL query support was explicitly requested.

5. **User interaction and exploration.** User interaction requirements necessitated both directed search capabilities, such as faceted queries for artists and artworks, and mechanisms for open-ended exploration. The interface was required to support visual navigation of the network and remain accessible to users regardless of their prior domain knowledge, with particular attention to those lacking SPARQL expertise.

6. **Visualization.** Visualization requirements, as articulated by domain experts, included the provision of graph-based representations of co-occurrences and relationships, facilities for visual exploration of emergent patterns, and views that make explicit instances in which artworks manifest collaborative effects, such as shared authorship.

7. **Mention typology and entity resolution.** The Vasarian text references entities through multiple mechanisms, shown in Figure 7.17, proper names such as Michelagnolo, definite descriptions such as *the statue*, and coreferential expressions such as *it*. Consequently, the ontology must treat explicit, implicit, and coreferent mentions as first-class objects and systematically associate each mention with its canonical referent. This structure enables downstream queries to aggregate across mentions or to target specific surface forms. The typed mention–entity distinction is thus not merely a technical convenience but a methodological necessity underpinning the co-occurrence and artwork clusters discussed above.

8. **Extraction provenance and positional anchoring.** Since mentions, entities, and their links are generated by an LLM rather than directly extracted from the text, the ontology must systematically record the conditions under which they are produced: the specific model employed and the prompt template under which it was invoked. Each mention must be anchored to its precise textual locus, at minimum the containing paragraph, and, when greater granularity is required, a character offset. This ensures that every extracted

statement remains traceable to the passage that warrants it. The resulting provenance layer enables the KG to serve as a documented interpretation rather than an unattributed assertion. A scalar per-mention confidence value was envisaged in the original specification as a complementary signal alongside the mention typology, but is deferred to future work (Section 11.9).

Domain experts underscored the need to distinguish versions of the work in the bibliographic layer – differentiating work-, expression-, and manifestation-level metadata and modeling dependencies between content and its sources. This follows directly from conducting all extraction on an English translation rather than Vasari's 16th-century Italian, since any inquiry into Vasari's Renaissance must trace extracted entities to the precise edition and translation, linking each to its paragraph in the translation with persistent references to web-accessible versions. The KG thereby makes explicit that its content reflects de Vere's early-twentieth-century interpretation of Vasari, not direct access to the sixteenth-century Italian source, and anticipates future comparison across translations or with the original.

This requirement exceeds the scope of the basic bibliographic metadata initially assumed in Section 7.3.1, demanding comprehensive transparency regarding the provenance of the source text. The ontology must record not only the extraction of a co-occurrence but also its exact linguistic and editorial context within Vasari's corpus. For domain experts, bibliographic representation must operate at the level of individual web pages. Merely indicating that a co-occurrence exists somewhere in de Vere's translation is insufficient; the ontology must provide direct, persistent links to the specific web page containing the relevant paragraph in an online edition, such as the Project Gutenberg HTML version.

The requirements mandate distinct ontological classes for the abstract work, for each individual edition – such as the 1568 Italian edition and modern critical editions – and for translations, including de Vere's 1912 version and other English or German ones. The ontology must further differentiate manifestations at multiple levels of granularity, modeling volumes, biographies, pages, and paragraphs as discrete, web-accessible entities with persistent URLs. For each extracted data element, provenance must be recorded with precision, linking it to the specific paragraph, page, volume, and translation from which it originates. Consequently, the ontology must enable the simultaneous representation of multiple editions and translations.

As previously established, extending existing best practices is essential for interoperability. Many requirements are already addressed by these practices; for example, FaBiO (Peroni and Shotton 2018) provides the structural foundation for the bibliographic module, which in turn motivated the introduction of web representation classes during the second design iteration (see Section 7.3.2.2).

7.3.2.2 Viewsari phase II.2: Improving the ontology module

The ontology in the prior sections is a preliminary implementation built from the initial data requirements and user interviews. Adopting XD enabled the subsequent requirement expansion (Section 7.3.1.2) through iterative modeling and systematic validation with domain experts and emergent data. We now formalize these expanded requirements in the new modules.

Improving the bibliographic layer. The initial ontology introduced a lightweight work-level module using the classes `:Book`, `:Edition`, `:Translation`, and `:Volume`, which were inspired by the FRBR and FaBiO structure but not a direct extension of them (IFLA Study Group on the Functional Requirements for Bibliographic Records 2009, Peroni and Shotton 2018) (the bibliographic-vocabulary landscape from which both descend is surveyed in Section 5.2.2.3 and Section 5.2.2.3). While this was sufficient for the first iteration, the bibliographic requirements identified in Section 7.3.2.1, including editions, translations, volumes, and the need to link directly to online source-text passages, necessitate an existing ontology as the foundation. Accordingly, we adopted FaBiO as the structural backbone because it provides a well-established four-level architecture that extends the concepts proposed by FRBR, namely Work, Expression, Manifestation, and Item. Concretely, the requirement arose to link instances in the KG not only to the abstract work and its editions or translations but also to *specific web-accessible manifestations* of those editions, e.g., the Project Gutenberg version that we used for entity extraction. This requirement is directly tied to the project's commitment to source text transparency. As discussed before, domain experts emphasized that working with an English translation as the sole source text necessitates *explicit clarity* about what text is being utilized.

Figure 7.18 presents the bibliographic layer. This component was architecturally foundational and maintained structural stability throughout successive iterations. In response to evolving requirements, the bibliographic layer was systematically redesigned to incorporate explicit representations of the Work, Expression, and Manifestation levels. The introduction of the classes `viewsari:edition`, `viewsari:translation`, `viewsari:volume`, `viewsari:biography`, and `viewsari:page` aligns these entities with `fabio:expression`, thereby inheriting established structural relationships while preserving Viewsari-specific semantics. To capture granularity of manifestations, the ontology now defines `viewsari:volume web representation`, `viewsari:biography web representation`, and `viewsari:page web representation` as subclasses of `fabio:WebPage`. This replaces the undifferentiated manifestation node of the initial design with a structured web layer that directly reflects the document hierarchy. Specifically, `viewsari:volume web representation` denotes the web representation of an entire volume, such as the HTML page containing all of Volume 3 on Project Gutenberg (hence it is a subclass of `fabio:WebSite`). `viewsari:biography web representation` identifies the

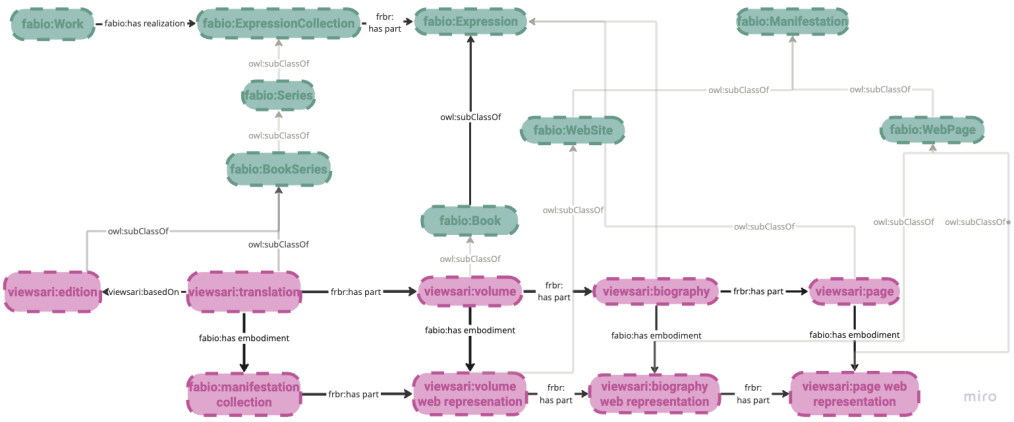


Figure 7.18: The Viewsari bibliographic layer and its alignment with FaBiO and FRBR. Green dashed nodes represent classes reused from FaBiO and FRBR (`fabio:Work`, `fabio:Expression`, `fabio:Manifestation`, `fabio:Book`, `fabio:WebPage`, and related collection classes); pink dashed nodes represent Viewsari-defined subclasses that specialize these for Vasari's *Lives*, including `viewsari:translation` (a `fabio:Book`), the part-of hierarchy `viewsari:volume` → `viewsari:biography` → `viewsari:page`, and their HTML counterparts at the manifestation level. Solid black arrows denote FaBiO and FRBR object properties (`fabio:hasRealization`, `frbr:hasPart`, `fabio:hasEmbodiment`); dotted grey arrows denote `owl:subClassOf` relationships.

web page or section corresponding to a single artist biography within a volume. In cases where online editions provide individual pages for each biography, as exemplified by Project Gutenberg, this class enables precise linking to specific biographies, such as Michelangelo's. Finally, `viewsari:page web representation` models individual pages within a web-accessible volume or biography. This is essential for paginated online editions, where each page is associated with a stable URL or fragment identifier, enabling integration with the corresponding biography structure. These classes are linked to their corresponding abstract textual units (the Expressions) via `fabio:isEmbodimentOf`. This design choice acknowledges that "a volume" constitutes an intellectual construct, distinct from "the HTML file hosting that volume on Gutenberg," which is a digital artifact characterized by technical properties such as URL, encoding, and hosting infrastructure. Introducing web manifestation classes directly addresses domain experts' concern about using an English translation as the exclusive source text and makes it very clear where the information was extracted from, as it is accessible via clickable links to the exact web page hosting each paragraph. For example, a researcher querying the KG might ask:

> Show me all co-occurrences between Leonardo da Vinci and Verrocchio, along with the source paragraphs and their web URLs.

Moreover, the web manifestation layer supports comparative work across editions and translations. If the KG later incorporates co-occurrences extracted from the 1568 Italian edition, users can query:

> For co-occurrences involving Michelangelo, compare the English translation's web manifestation with that of the Italian edition.

An example of the concrete instantiation of this structure is illustrated in Figure 7.19, which traces Vasari's *Le Vite* from the abstract `fabio:Work` through the 1568 edition and de Vere translation down to volume, biography, and page-level embodiments, alongside a parallel `fabio:WebManifestation` branch for the Project Gutenberg version. We elaborate on the concrete instantiation of the structural and content layers and how they attach to this bibliographic backbone in Section 7.3.2.2 and Section 7.3.2.2.

The introduction of web manifestation classes had cascading effects on the ontology design, as provenance chains became longer. Instead of a co-occurrence linking directly to a paragraph, which links to a volume, which links to a translation, the chain now includes intermediate web-representation nodes (visible in Figure 7.19 as the lower row of `viewsari:*_web_representation` instances) and provenance information about activities and agents. This increases the complexity of SPARQL queries but ensures that every link in the chain carries the appropriate metadata. This can be overcome using shortcuts in the ontology, as we explain in Section 7.3.2.2.

Structural layer: Document hierarchy and paragraphs. The structural layer addresses the requirement for positional provenance, enabling the extracted entities to be linked to the specific paragraph(s) from which they originate. This module draws on the DoCO ontology (Document Components Ontology) (Peroni and Shotton 2018) to represent discourse elements such as biographies and paragraphs, with `doco:Paragraph` and `doco:TextChunk` serving as a structural anchor carrying the properties needed to locate a passage within the document hierarchy as well as the text snippet itself. Figure 7.20 illustrates this instantiation for paragraph 103 of volume 1, showing how the paragraph's textual content and its embodiment as a `doco:TextChunk` on the Gutenberg web page jointly support the anchoring of entity mentions.

The structural layer serves as the point of attachment for the content layer. Extracted entities and co-occurrences, described in detail in Section 7.3.2.2, are instantiated as `oa:Annotation` objects according to the Web Annotation Ontology (Sanderson et al. 2017). The mechanisms responsible for anchoring these annotations to the text, specifically the `doco:TextChunk` targets and `oa:TextPositionSelector` selectors with their associated character offsets, are integral components of the structural layer. In this architecture, the annotation node functions as an

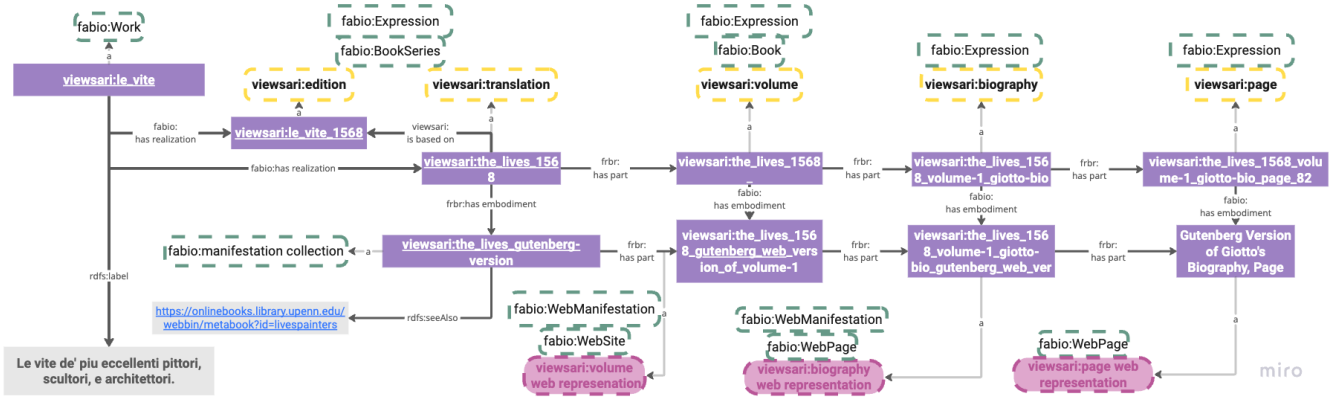


Figure 7.19: Bibliographic-layer example: Vasari's *Le Vite* modeled with FRBR/FaBiO, from the abstract `fabio:Work` through the 1568 edition and de Vere translation down to volume, biography, and page-level embodiments, with the Project Gutenberg web manifestation as a parallel realization.

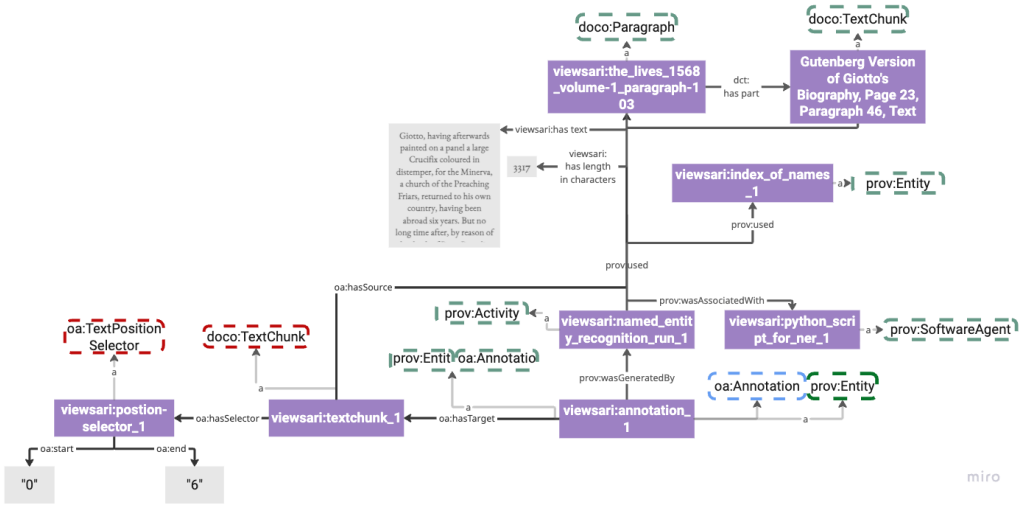


Figure 7.20: Structural-layer example: paragraph 103 of volume 1, with its plain-text content, character length, and embodiment as a `doco:TextChunk` on the Gutenberg web page. A named-entity annotation anchors the mention "Giotto" to the chunk via a `oa:TextPositionSelector`, with PROV-O recording the extraction activity, the extraction agent, and the index of names used as input.

interface between the two layers: the *body*, representing the extracted mention or relation, is assigned to the content layer, whereas the *target* and *selector* are classified as structural elements. This separation ensures that bibliographic and textual structures remain reusable across multiple extraction runs and content types. As a result, identical paragraph and chunk instances can accommodate any number of annotations without redundant positional metadata.

At the structural level, `viewsari:has start page` (domain: `doco:Paragraph`, range: `viewsari:page web representation`) and `viewsari:has end page` (same domain and range) record the opening and closing printed page of each paragraph as direct links to the corresponding page web representation node. This reflects what is observed in the corpus, since 48.2 % (1,676 of 3,479 paragraphs) of the de Vere translation span more than one printed page (Table 6.3); a single-page-link property cannot represent this without information loss, making these two properties a structural necessity. The initial property `viewsari:on`, which linked a paragraph to a single page, is deprecated in the current version and replaced by this pair.

Content layer: the extracted-content superclass. The initial ontology defined domain-specific content classes, such as `viewsari:person` and `viewsari:artwork`, but omitted a unifying superclass. As the extraction pipeline incorporated additional entity types generated by heterogeneous methods, including statistical co-occurrence analysis and LLM-guided extraction,

this lack of a shared abstraction introduced a significant modeling constraint. In particular, the absence of a common superclass complicated the formulation of generic SPARQL queries across entity types, necessitating UNION clauses. This limitation underscores the need for an ontological structure that supports both methodological extensibility and consistent representation of provenance.

This requirement led to the introduction of `viewsari:extracted content` as the superclass of all content-layer individuals, each of which is also typed as `prov:Entity`. This dual typing is a principled design choice. Every extracted resource, whether a person identified in the Index of Names or an artwork recognized by ObliquER, is simultaneously a domain entity and the documented result of a computational extraction process. This structure ensures that provenance is systematically attached to all content-layer entities, supporting both interpretability and reproducibility. The closest vocabulary-level analog to this commitment is HiCO's modeling of historical claims as first-class, source-attributed assertions (Section 5.2.2.4); Viewsari diverges from HiCO in attaching the same provenance discipline to *every* extracted content node rather than only to interpretive claims, on the grounds that the boundary between description and interpretation is itself one of the things the graph is meant to render visible. The elevation of `viewsari:cooccurrence` to this hierarchy represents a substantive extension to the ontology, meaning that the status of the co-occurrence as an observable entity from the extraction process becomes explicit. Previously, `viewsari:cooccurrence` was defined solely as a domain-specific class, lacking inheritance from `prov:Entity`. The current version rectifies this limitation, ensuring that co-occurrences, like all other entities, systematically carry extraction provenance.

Figure 7.21 and Figure 7.22 show the dual-typing pattern in operation for two artwork mentions extracted from the de Vere translation. In both examples, the mention node carries the joint type `oa:Annotation` ⊓ `prov:Entity` and the resolved artwork entity carries the joint type `viewsari:artwork` ⊓ `prov:Entity`. The annotation is anchored to its textual evidence via an `oa:hasTarget` link to a `doco:TextChunk` whose `oa:TextPositionSelector` specifies the character offsets in the source paragraph. For the annotation, the textual surface form is recorded using `oa:hasBody`, which connects it to a `oa:TextualBody` with `rdf:value` containing the plain text value. The resolved entity carries an `owl:sameAs` link to its Wikidata identifier. Each link in the chain – mention, entity, and the two pipeline runs (`prov:Activity`) that produced them – is reified, with `prov:wasGeneratedBy` edges connecting the artifacts to the typed activities and `prov:wasAssociatedWith` edges pointing back to the executing software agents.

Figure 7.21 illustrates the minimal instantiation of the pattern, namely an explicit artwork mention with a single, fully linkable referent and no further domain enrichment. Figure 7.22 reuses the same template for Michelangelo's *Pietà*, with an additional `wd:P170` (*creator*) edge connecting the artwork entity back to `viewsari:michelangelo`. Domain relations between extracted-content

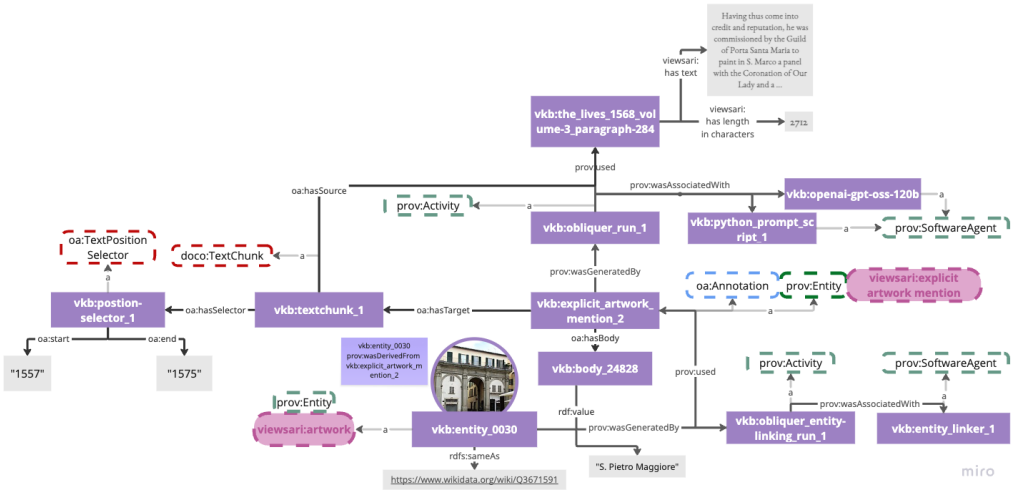


Figure 7.21: Worked example of the extracted-content layer for the explicit artwork mention *S. Pietro Maggiore* in paragraph 284 of volume 3 of the de Vere translation. The mention is dual-typed as `oa:Annotation` and `prov:Entity`; the linked artwork entity is dual-typed as `viewsari:artwork` and `prov:Entity`. The two ObliquER pipeline runs that produced the mention and its Wikidata linking are reified as typed `prov:Activity` instances and connected to their software agents.

entities therefore inhabit the same provenance graph as the mentions that surface them, rather than being layered on top as a separate annotation; the artist–artwork link is reachable through the same `prov:wasGeneratedBy` chain as the mention itself, and is itself an extracted-content artifact rather than a curated assertion.

The retyping of `viewsari:cooccurrence` constitutes a conceptual reframing rather than a mere inheritance adjustment. In the first design (Section 7.3.1.2), co-occurrence was modeled as an event according to the Participation pattern, where two persons participated in a shared situation, and the situation was reified as a first-class node. This approach introduces an ontological commitment unsupported by the evidence, as it implies a real-world relationship between the participants. In practice, two artists who co-occur in a Vasari paragraph may not have met, collaborated, or even been contemporaries. Their co-occurrence reflects only that Vasari, or the extractor interpreting Vasari, mentioned them within the same textual unit.

Modeling co-occurrence as an event would collapse precisely the mention/referent relation that our central claim does not support. The current design, therefore, types `viewsari:cooccurrence` as a subclass of `viewsari:extracted content` (and thus of `prov:Entity`) rather than of any event class, instantiating it directly with a `viewsari:involves` link to each participating entity and a shortcut `viewsari:in paragraph` link to its source. This preserves the provenance of

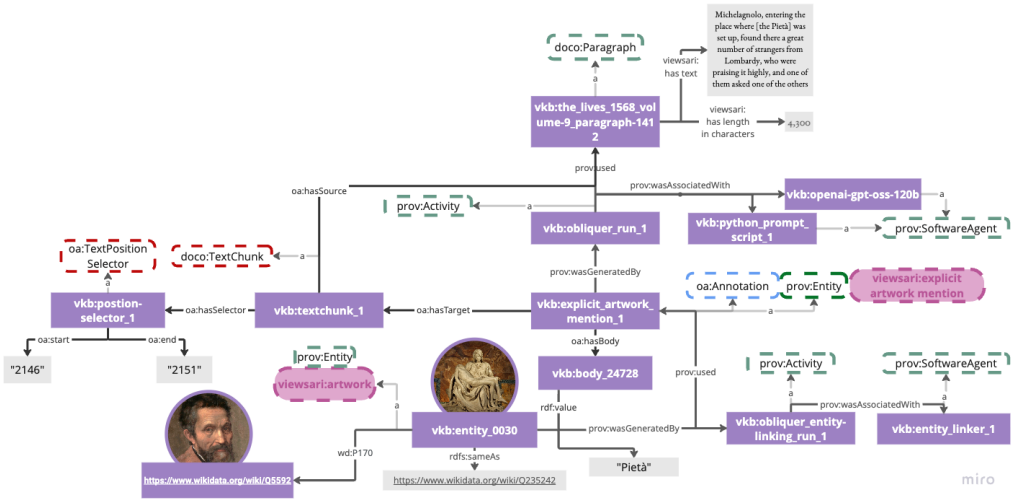


Figure 7.22: Worked example of the extracted-content layer for the artwork *Pietà* in paragraph 1412 of volume 9, with an additional `wd:P170` (*creator*) link from the artwork entity to `viewsari:michelangelo`. The dual-typing and provenance scaffolding are identical to Figure 7.21; only the domain enrichment differs.

the original first-class shape, but on a foundation that correctly reflects what a co-occurrence *is* in this work, namely a record of textual co-mention, not a record of an event.

The mention taxonomy: from a single annotation class to a typed hierarchy. The requirements identified during ObliquER development (see Section 7.4) demonstrate that mention type is a foundational element of the Viewsari KG rather than merely a metadata annotation on a generic resource. Distinguishing among explicit artwork mentions, descriptive implicit references, and coreferential pronouns is methodologically significant since these categories directly shape the epistemic evaluation of extraction methods and the validation of results against the GT.

This motivated the introduction of `viewsari:mention` as an intermediate class that consolidates the shared structure of all mention types:

$$\texttt{viewsari:mention} \sqsubseteq \texttt{oa:Annotation} \sqcap \texttt{prov:Entity}$$

The double superclass, comprising `oa:Annotation` and `prov:Entity`, formalizes that each mention functions simultaneously as a scholarly annotation, carrying `oa:hasTarget`, `oa:hasBody`, and `oa:hasSource`, and as a provenance artifact, carrying `prov:wasGeneratedBy` and `prov:wasDerivedFrom`.

From `viewsari:mention`, four typed subclasses are derived, each addressing a distinct mention phenomenon identified in the ObliquER annotation guidelines (Section 7.4.6.1):

`viewsari:explicit mention` A mention in which the entity is named through a canonical surface form or established title. Typed as $\sqsubseteq$ `viewsari:mention`.

`viewsari:implicit mention` A mention in which the entity is recoverable from the passage but not named through a canonical surface form. Vasari's descriptive circumlocutions, such as *"a panel he executed for the high altar"*, exemplify this phenomenon. Typed as $\sqsubseteq$ `viewsari:mention`.

`viewsari:coreferent` A mention that refers back to an entity already introduced in the same or a preceding paragraph, realized as a pronoun, demonstrative, or ellipsis. Typed as $\sqsubseteq$ `viewsari:mention`.

`viewsari:generic mention` A mention that refers to an entity type in general rather than to a specific individual; e.g., *"works of painting"* as a reference to artistic practice rather than to a specific artwork. Generic mentions are excluded from the computation of ObliquER precision and recall but are retained in the graph to support studies of Vasari's rhetorical register.

Two further specializations address the domain-specific needs of artwork extraction:

`viewsari:explicit artwork mention` defined as an explicit mention whose body is an `viewsari:artwork` individual.

`viewsari:implicit artwork mention` is defined as an implicit mention whose referent is an artwork recoverable from the passage by inference but not named explicitly.

Retaining `viewsari:coreferent` and `viewsari:generic mention` expresses a hermeneutic orientation toward representing the structurally significant properties of the source text, enabling the KG to encode them accordingly.

Activity classes: formalizing the pipeline as typed provenance. The initial provenance module introduced the Provenance pattern but did not fully implement agents, activities, and related features. As the pipeline grew to encompass not only NER and coreference resolution from the Index of Names or co-occurrence extraction, but also LLM-guided entity recognition and linking, it became necessary to formally type activity instances so that the KG can filter and reason over provenance chains by activity kind (and thus, adhere to the requirements given by the domain experts). The choice to specialize `prov:Activity` into a small taxonomy of

pipeline-specific subclasses, rather than to adopt CIDOC CRM's (Doerr 2003) event-centric pattern (Section 5.2.2.1) or to remain at the generic level of PROV-O (Section 5.2.2.2), is motivated by the requirement that downstream users be able to filter by extraction method without traversing a free-text activity label, while preserving full interoperability with PROV-O through the standard subclass hierarchy. The current version of the module, therefore, refines the generic `ProvenanceActivity` of the pattern into a small taxonomy of activity subclasses, one per pipeline stage, each carrying the parametrization relevant to its kind (model identifier and prompt version for the LLM-guided steps, reference for the Index of Names-based extraction, etc.). `viewsari:cooccurrence` instances and the entity- and mention-level resources produced upstream are linked via `prov:wasGeneratedBy` to the typed activity that produced them, and the activity in turn carries `prov:used` edges back to the paragraph-level resources. The `Agent` is realized as a software agent (the pipeline component, versioned).

The three new `prov:Activity` subclasses are, in detail:

`viewsari:named entity recognition` – represents a run of the NER and coreference resolution pipeline over the Index of Names. Instances carry `prov:used` links to the index-of-names entities and the involved paragraphs, and `prov:wasAssociatedWith` links to the executing software agent.

`viewsari:cooccurrence analysis` – represents a run of the co-occurrence extraction pipeline over the paragraph corpus. Instances record which paragraphs were consumed (`prov:used`) and which co-occurrence entities were generated (`prov:wasGeneratedBy`).

`viewsari:entity linking` – represents a run of the ObliquER EL stage, in which recognized mentions are matched against a Wikidata candidate set. Instances are associated with the LLM software agent and carry links to the prompt template version used, enabling reproducibility across pipeline runs.

Typing activities formally rather than relying on instance-level labels reflects the fact that each pipeline stage is not a single event but a class of events, the Index-based NER and coreference resolution, the co-occurrence extractor, and the ObliquER LLM-guided extraction are each executed several times – across different paragraph batches, prompt revisions, model versions, and re-runs over the corpus – such that the KG accumulates a population of activity instances per stage. Treating the stages as classes lets each individual run be instantiated as its own `prov:Activity` with its own timestamp, parametrization, and inputs, while remaining recognizable as a member of the same class.

An example for this is CQ CQII.39 from Section A.0.0.0.2:

For a given extraction result, can the provenance chain be traced back from the entity through the annotation, the paragraph, the page, the volume, and the translation to the original work?

Refinement of the properties. At the work level, persons are assigned to a `fabio:Work` via the `dct:creator` property, establishing authorship relationships.

From the work level to the instantiation level, the following properties establish relationships between works and their physical or digital embodiments. `viewsari:is based on` (⊑ `frbr:relatedEndeavour`, domain: `viewsari:translation`, range: `viewsari:edition`) records the derivation relationship between the de Vere translation and the 1568 edition. This property replaces an informal `rdfs:seeAlso` link used in the initial prototype with a typed FRBR relation that supports reasoning over the bibliographic hierarchy. `frbr:hasPart` connects volumes to their containing translation, and `frbr:hasEmbodiment` links expression-level instances to their manifestation-level counterparts; both are inherited from FaBiO and FRBR rather than redefined in Viewsari.

At the structural level, `viewsari:has start page` and `viewsari:has end page` (domain `doco:Paragraph`, range `viewsari:page web representation`) replace the deprecated single-page `viewsari:on`, recording paragraph extent across the 48.2 % of paragraphs that span multiple printed pages (Section 7.3.2.2).

Two data type properties complete the structural layer. `viewsari:has text` (⊑ `oa:bodyValue`) stores the full normalized text of a `doco:Paragraph` instance as a plain string literal.

Across the refinements introduced above, every Viewsari-defined property is asserted as a subproperty of an established external term: `dct:relation` subsumes the content-layer anchoring properties `viewsari:in paragraph` and `viewsari:involves`; `frbr:relatedEndeavour` subsumes `viewsari:is based on`; `oa:bodyValue` subsumes `viewsari:has text`.

Fixed individuals: anchoring bibliographic nodes. The current version of the ontology includes four named individuals that serve as the stable root nodes of the bibliographic layer. Their introduction resolves a modeling gap in the first version, where the top of the hierarchy was represented only at the class level, with no instances populated.

`viewsari:le_vite` **(`#0001027`)** An individual of type `fabio:Work` representing the abstract intellectual work *Le Vite de' più eccellenti pittori, scultori, e architettori*, linked to its Wikidata counterpart (`wd:Q1645493`) via `rdfs:seeAlso`.

`viewsari:le_vite_1568` **(#0001028)** An individual of type `viewsari:edition`, representing the canonical 1568 Giunti edition as the primary expression of the work in the Viewsari ontology.

`viewsari:the_lives_1568` **(#0001029)** An individual of type `viewsari:translation`, representing the 1912 de Vere English translation. This is the expression from which all volume, biography, and page instances in the current KG descend via `frbr:hasPart`.

`viewsari:the_lives_gutenberg_version` **(#0001031)** It represents the Project Gutenberg web publication of the de Vere translation in its entirety and is of type `fabio:ManifestationCollection`. All ten `viewsari:volumewebrepresentation` instances are linked to this node via `frbr:isPartOf`. An `rdfs:seeAlso` link points to the Penn Online Books meta-record for the edition.

Together, these four individuals instantiate the complete bibliographic spine of the Viewsari KG, running from the abstract Vasarian work, through the 1568 Italian edition and the 1912 English translation, to the Project Gutenberg web collection, with every volume, biography, page, and paragraph hanging from this spine via typed part-of and embodiment relations.

7.3.2.3 Summary of changes and their motivation

The Phase II refinements respond to motivating shifts that emerged after the first ontology version entered productive use. The realization that LLMs function as interpretive agents (Section 7.3.2.1) made the distinction between textual mention and real-world referent representationally load-bearing rather than incidental; every modeling decision in this iteration has to make that distinction explicit, since collapsing it is precisely what the rejected PIEA (C1) does. The move to a generative-extraction pipeline with versioned prompt templates, model identifiers, and reasoning traces made provenance a constitutive rather than ancillary feature of the ontology, demanding typed activity classes in place of a generic provenance pattern (C2). Domain experts' insistence on bibliographic transparency – that every extracted assertion remain traceable to the precise edition, translation, and web-accessible source page – demanded a properly stratified bibliographic layer aligned with established CH vocabularies rather than the ad hoc class set introduced in the first ontology version.

These motivations drove changes across all three layers of the architecture and the property surface, organized as follows.

Bibliographic layer. The first ontology version features classes like `viewsari:book`, `viewsari:edition`, `viewsari:translation`, and `viewsari:volume`, which are realigned with

FaBiO and FRBR, so that the Work / Expression / Manifestation distinction is now structurally enforced rather than merely suggested. Three new web-representation classes (`viewsari:volumewebrepresentation`, `viewsari:biography webrepresentation`, `viewsari:pagewebrepresentation`) extend the manifestation layer with referenceable links to the Project Gutenberg digital edition, satisfying the requirement that every co-occurrence and mention link back to a specific online paragraph. Four fixed individuals (`viewsari:le_vite`, `viewsari:le_vite_1568`, `viewsari:the_lives_1568`, `viewsari:the_lives_gutenberg_version`) anchor the bibliographic spine at the instance level and replace the first ontology version class-only top of the hierarchy.

Structural layer. `doco:Paragraph` replaces the first ontology version `viewsari:paragraph` class, and paragraph extent is now formalized through the pair `viewsari:has start page` and `viewsari:has end page` pointing to `viewsari:page web representation` instances – a refinement that accommodates the empirical fact that 48.2 % of paragraphs in the de Vere translation cross a printed page boundary (Table 6.3), which the first single-page property could not represent without loss. Two datatype properties, `viewsari:has text` (subsuming `oa:bodyValue`) and `viewsari:has length in characters`, complete the structural layer for selector validation and corpus statistics.

Content layer. The most substantial refinement introduces `viewsari:extracted content` as a uniform superclass for every content-layer individual, dual-typed as `prov:Entity`. This dual typing makes provenance a structural feature of every domain entity rather than an external annotation. `viewsari:cooccurrence` is retyped as a subclass of `viewsari:extracted content` rather than of any event class, correctly reflecting its status as a textual co-mention rather than a real-world event. The mention taxonomy – `viewsari:mention` subsuming `viewsari:explicit mention`, `viewsari:implicit mention`, `viewsari:coreferent`, and `viewsari:generic mention`, with two further artwork-specific specializations – turns mention type into a first-class queryable feature of the graph. `viewsari:organization` captures institutions and guilds in Vasari's network. At the data level, ambiguity preservation, with pipe-separated identities in the Index of Names instantiated as separate `viewsari:person` individuals rather than collapsed into one, operationalizes the rejection of the PIEA in the populated graph (C1).

Activity classes and properties. The generic provenance pattern of the previous ontology version is refined into three typed activity classes corresponding to pipeline stages: `viewsari:named entity recognition` for the Index-of-Names recognition and coreference

resolution, `viewsari:cooccurrence analysis` for the pointwise-mutual-information extractor, and `viewsari:entity linking` for the ObliquER LLM-guided linking stage. Each activity instance carries the parametrization relevant to its kind (model identifier, prompt-template version, executing software agent) and supports typed traversal of provenance chains by activity kind (**C2**). On the property surface, `viewsari:is based on` formalizes the typed FRBR relation between translation and source edition; `viewsari:in paragraph` and `viewsari:involves` anchor co-occurrences to source paragraphs and participating entities under `dct:relation`. Every Viewsari-defined property is asserted as a subproperty of an established external term, so the namespace introduces no orphan vocabulary above the upper-level vocabulary on which the refined ontology builds.

Table 7.3 provides an overview of the external vocabularies reused as the backbone of the three-layer architecture – FaBiO, FRBR, and DoCO for the bibliographic and document-structural layers, PROV-O and the W3C Web Annotation Ontology for provenance and annotation, and Dublin Core Terms for the generic relation and authorship subproperties – together with the specific classes and properties used from each. Table 7.4 provides the complementary inventory of Viewsari-defined elements introduced in this iteration, listing the motivating requirement and the pipeline stage that triggered each addition. Together, these two tables form the complete vocabulary surface of the refined ontology.

Toward two pattern contributions. The first-class textual-observation shape of `viewsari:cooccurrence` introduced above, together with the typed-activity and prompt-recording extensions of the Provenance pattern noted earlier in this subsubsection, emerge only once the extraction layer is in place. They constitute the basis for two candidate pattern contributions of this work, the formal proposals for which are deferred to Section 10.3.2 and Section 10.3.3: the *Co-occurrence* pattern, generalizing `viewsari:cooccurrence` into a domain-neutral template for evidence-anchored textual or computational observations, and *Computational Provenance* (COMP-PROV), which abstracts the activity-typing and prompt-recording commitments into a reusable specialization of the Provenance pattern.

This supports both reproducibility of results and critical interpretation, since a co-occurrence derived from point-wise mutual information carries a different epistemic weight than an implicit artwork mention inferred by an LLM from a descriptive passage, and the ontology must formalize this distinction.

These revisions exemplify a recurrent dynamic in ontology development within the DH, in which initial modeling choices, while sufficient for narrowly defined objectives, inevitably become constraints as the project scope expands. The iterative methodology of XD enables systematic identification and resolution of such limitations without necessitating a wholesale redesign. Rather

Table 7.3: External vocabularies reused by the Viewsari ontology, with the specific classes and properties used from each. Viewsari-defined terms inherit from these vocabularies via `rdfs:subClassOf` and `rdfs:subPropertyOf` axioms, so the ontology introduces no orphan vocabulary above the level of these imports.

Prefix	Vocabulary	Classes reused	Properties reused
`fabio`	FaBiO (Peroni and Shotton 2018)	`Work`, `Expression`, `Manifestation`, `WebPage`, `WebSite`, `ManifestationCollection`	`isEmbodimentOf`
`frbr`	FRBR core (IFLA Study Group on the Functional Requirements for Bibliographic Records 2009)	–	`hasPart`, `hasEmbodiment`, `relatedEndeavour`, `isPartOf`, `realizationOf`
`doco`	DoCO (Peroni and Shotton 2018)	`Paragraph`, `TextChunk`	–
`prov`	PROV-O (Lebo et al. 2013)	`Entity`, `Activity`, `Agent`, `SoftwareAgent`	`wasGeneratedBy`, `wasDerivedFrom`, `wasAssociatedWith`, `used`
`oa`	W3C Web Annotation Ontology (Sanderson et al. 2017)	`Annotation`, `TextPositionSelector`, `TextualBody`	`hasTarget`, `hasBody`, `hasSource`, `hasSelector`, `start`, `end`
`dct`	Dublin Core Terms (DCMI Usage Board 2020)	–	`creator`, `relation`

than producing a fundamentally different ontology, this process yields a refined model that preserves the original ontological commitments while enhancing precision and interoperability. Consequently, the Viewsari KG functions not merely as a repository of extracted facts but as a transparent record of interpretive processes, substantiating the methodology's hermeneutic claims through machine-readable, queryable provenance.

The applicability and quality of the formalizations mentioned above are evaluated using CQs and domain-expert feedback; we describe this in detail in Chapter 9.

7.4 The ObliquER Pipeline

ObliquER serves as the empirical realization of this dissertation's central claim that KE from interpretive sources is itself an interpretive act, necessitating systematic provenance documentation.

Table 7.4: New ontology elements introduced in the current version relative to the first formalization, with the motivating requirement and the pipeline stage that triggered each addition.

Element	Type	Motivating requirement
`viewsari:extracted content`	Class	Uniform provenance superclass across all entity types
`viewsari:organization`	Class	Institutions and guilds in Vasari's network
`viewsari:mention`	Class	Shared dual OA/PROV structure for all mentions
`viewsari:explicit mention`	Class	ObliquER mention taxonomy
`viewsari:implicit mention`	Class	ObliquER mention taxonomy
`viewsari:coreferent`	Class	Coreference chains across paragraphs
`viewsari:generic mention`	Class	Non-entity discourse references
`viewsari:explicit artwork mention`	Class	Fine-grained artwork typing
`viewsari:implicit artwork mention`	Class	Fine-grained artwork typing
`viewsari:named entity recognition`	Class	Typed NER activity provenance
`viewsari:cooccurrence analysis`	Class	Typed co-occurrence activity provenance
`viewsari:entity linking`	Class	Typed EL activity provenance
`viewsari:has start page`	Obj. prop.	Cross-page paragraph extent
`viewsari:has end page`	Obj. prop.	Cross-page paragraph extent
`viewsari:in paragraph`	Obj. prop.	Scoped co-occurrence anchoring
`viewsari:involves`	Obj. prop.	Generalized co-occurrence participation
`viewsari:is based on`	Obj. prop.	Typed translation–edition relation
`viewsari:has text`	Data prop.	Full paragraph text storage
`viewsari:has length in characters`	Data prop.	Corpus statistics and selector validation
`viewsari:le_vite` (#0001027)	Individual	Bibliographic root: abstract work
`viewsari:le_vite_1568` (#0001028)	Individual	Bibliographic root: 1568 edition
`viewsari:the_lives_1568` (#0001029)	Individual	Bibliographic root: de Vere translation
`viewsari:the_lives_gutenberg_version` (#0001031)	Individual	Bibliographic root: Gutenberg manifestation collection

In direct response to the provenance gap identified in C2 and to operationalizing C3 at the system level, ObliquER reframes entity recognition and linking not as neutral surface-form detection but as an interpretive process whose epistemic context must be recorded. Each extraction decision is anchored to the specific extraction activity, the prompt template, and the model version responsible for its production. This also means that we explicitly differentiate between the detection of mentions (their textual forms) and the actual referents, which are constructed from the latter using EL. Detected mentions are clustered according to shared entity identifiers, ensuring consistent referential integrity. The pipeline leverages LLMs with prompt templates grounded in the Viewsari ontology, thereby formalizing the recovery of implicit information from textual sources and its integration into a provenance-aware KG. Both entity recognition and EL are addressed within a unified pipeline that treats explicit and implicit entity mentions as first-class objects. Each recognized entity mention is linked to its corresponding entry in an external KB, where available; otherwise, it is designated as OOKB and remains referenceable within the Viewsari KG.

ObliquER builds directly on the state-of-the-art literature reviewed in Chapter 5. The project draws its primary intellectual foundation from research on implicit entity recognition, as detailed in Section 5.1.4. This body of work demonstrates that descriptive, deictic, and other non-canonical references to entities constitute a distinct task class beyond canonical-mention NER. ObliquER advances this tradition by designating implicit mentions as a first-class output type, explicitly typed within the ontology and preserved throughout the graph population process (Section 7.4.2). The pipeline employs LLMs for both recognition and linking, thereby aligning with the generative IE paradigm (Section 5.1.2). Prompt-engineering strategies are systematically evaluated against the prompting-mode taxonomy introduced in Section 5.1.3. The EL component incorporates methods from generative EL research (Section 5.1.7) and addresses challenges identified in the long-tail and OOKB literature (Section 5.1.8), which inform the candidate generation and ranking procedures described in Section 7.4.9. Throughout this chapter, inline cross-references explicitly connect each design decision to its motivating prior work.

Author contributions. ObliquER, as a full pipeline, is the result of a collaborative effort with Cedric Möller, Jörg Waitelonis, and Harald Sack, as part of joint work currently under review. The EL components presented in Section 7.4.3 and candidate ranking were developed primarily in collaboration with Cedric Möller, who also served as an auxiliary annotator for the GT described in Section 7.4.4.

Implementation. ObliquER is released as an open-source repository.[16] The two pipeline stages, entity recognition (Section 7.4.8) and EL (Section 7.4.9), share the architecture of prompts, which are assembled from modular parts via a templating engine, inference works through an OpenAI-compatible client, and the dataset, model, and prompt-variant choices are composed at the command line through a Hydra configuration tree. Optional provenance logging records every prompt-rendering and inference event with a timestamp, model identifier, prompt identifier, and status, and these records populate the activity-instance metadata of the *NER* and *EL* activities described in Section 7.3.2.2.

Inference models. The primary inference model across both pipeline stages is `openai/gpt-oss-120b` (OpenAI 2025), served via an OpenAI-compatible endpoint – an on-premise inference engine server) for recognition runs linking runs, the model identifier is identical in both contexts. Decoding temperature is fixed at 0 to reduce output variance and improve reproducibility, and all other generation parameters are left at their defaults. The cross-model evaluation of Section 9.4.5 additionally runs the best-performing recognition prompt under `google/gemma-4-31b-it` (Gemma Team, Google 2026) to isolate the contribution of the decoder while holding the prompt fixed. The linking stage further relies on three pretrained candidate retrievers, used as recall backstops alongside the artist-constrained Wikidata pool: ReFinED (Ayoola et al. 2022), run end-to-end over each document to link explicit human mentions and to seed the artist-constrained artwork pool via `wdt:P170`; BLINK (Wu et al. 2020); and GENRE (De Cao et al. 2021), both invoked at the mention level for canonically named entities. As the core retriever, we use a single dense bi-encoder, fine-tuned on `Snowflake/snowflake-arctic-embed-m-v2.0` (Merrick et al. 2024), in two rounds, with the SemArt (Garcia and Vogiatzis 2018) and ArtPedia (Stefanini et al. 2019) GTs providing the training pairs after LLM-based metadata extraction. A `Qwen/Qwen3-Embedding-4B` model is used zero-shot during candidate-generation utility encoding as a strong off-the-shelf reference point against which the fine-tuned bi-encoder is compared in Section 9.4.4.

Prompt variants. The recognition stage supports three prompt variants. The variants are analyzed in detail in Section 7.4.8.4. For entity recognition, each variant uses two few-shot examples, where each example comprises a full paragraph. They are drawn from a small held-out subset of the corpus, excluded from evaluation. The linking stage parallels this with a standard disambiguation prompt, as described in Section 7.4.9.1. A development-split mechanism in the

[16] Project repository: `https://github.com/ISE-FIZKarlsruhe/ObliquER`. A complete inventory of scripts, configuration files, data directories, the dependency manifest, and per-component READMEs documenting their use is provided in the repository.

dataset configuration ensures that any paragraph used as a few-shot example is ignored from evaluation.

On ObliquER in the case of Vasari's *The Lives*. As a primary use case and proof of concept, we apply ObliquER to Giorgio Vasari's *The Lives*, focusing on extracting artwork entities whose mention types range from canonical titles to more intricate scene descriptions (Vasari 1568, 2008). Within the broader Viewsari project, ObliquER serves as the component for extraction and linking; its outputs, entity mentions, and their provenance-annotated links form the basis for populating the Viewsari KG based on the Viewsari ontology for a structured and epistemologically transparent representation (see Section 7.3).

This section provides a comprehensive description of the ObliquER methodology. Section 7.4.1 presents the motivation for ObliquER and the conceptual rationale for its name with Section 7.4.2 situating ObliquER within the Viewsari ontology and the requirements engineering process. Section 7.4.3 offers a formal definition of the recognition and linking tasks. Section 7.4.4.2 describes the corpus and the GT construction strategy, including the rationale for text sampling and the criteria for biography selection. Section 7.4.6 presents the annotation guidelines governing the GT annotation. Section 7.4.8 details the prompting strategy for entity recognition, while Section 7.4.9 outlines the EL strategy. Results, their evaluation, and a discussion of limitations are provided in Chapter 8 and Chapter 9.

7.4.1 Overview, motivation, and challenges

The Viewsari ontology, designed with art-historical domain experts, provides a formal structure for the knowledge embedded in *The Lives*. Together with ObliquER, it supports the central hypothesis of this dissertation, that systematic provenance tracking is essential for extraction pipelines, particularly as GenAI increasingly emulates human interpretive processes and presents interpretive choices as factual statements. The extraction of artwork entities emerged as a primary requirement, repeatedly emphasized by domain experts, given the foundational status of *The Lives* in art history. Unlike canonical proper nouns such as persons, locations, or organizations, artworks and related entities are often referenced implicitly, complicating their extraction. While prior work demonstrates that person entities can be reliably extracted using the Index of Names, identifying artworks, motifs, and other unconventional types remains a significant challenge due to their implicit nature.

To address these requirements, we introduce ObliquER, a framework leveraging LLMs to recognize and link both explicit and implicit entity mentions in art-historical texts. The framework is

designed to systematically address the phenomenon of implicitness and its epistemological implications within GenAI applications. Artworks serve as the primary use case, as they frequently lack canonical surface forms and are referenced only through descriptive passages.

The name ObliquER derives from the word oblique, "implied" or "hidden" and the acronym for 'entity recognition', and is a pun hinting that implicit references are more oblique (obliquer) and more difficult to recover than their explicit counterparts. This is a central characteristic of Giorgio Vasari's *The Lives*, in which entities such as artworks are often referred to by descriptive passages, location-based descriptions, or artist attributions rather than by formal titles.

ObliquER serves as the primary KE component within the broader Viewsari pipeline. Its outputs consist of structured documents summarizing recognized entity mentions, annotated with surface form, mention type, coreference cluster, and, where applicable, a Wikidata link to directly populate the Viewsari KG. The connection to the ontology is explicit, as ObliquER's outputs correspond to annotation classes anchored in the Viewsari ontology via the Web Annotation Ontology (Sanderson et al. 2017) and PROV-O (Lebo et al. 2013), ensuring that all extracted content remains traceable to its textual provenance.

Each recognized mention is represented in the Viewsari ontology. This design addresses, among others, CQs `CQII.6` (*Which artworks are explicitly named in the biography of artist X, and which are described only implicitly?*) and `CQII.7` (*What is the surface form by which artwork A is referred to in paragraph P, and is it an explicit, implicit, or coreferential mention?*) and related queries from the requirements engineering process detailed in Section 7.3.1 and Section 7.3.2.

These requirements arise from concrete challenges in applying NER and EL to Vasari's text, which persistently recur across both pipeline stages.

Absence of canonical titles. A substantial fraction of the artworks discussed by Vasari predate the establishment of systematic titling conventions (Welchman 1997, Yeazell 2015). Consequently, many works lack canonical titles and are referenced in art-historical literature through descriptive phrases, artist attributions, or by location and commissioner. Indeed, even contemporary cataloging standards in the CH domain acknowledge that a "title" is not necessarily a canonical, fixed reference but may take the form of an extended descriptive phrase, an iconographic identification, or an implicit designation by object type, dedication, or location. EL systems that depend on surface-form similarity to knowledge-base labels cannot reliably resolve such references, as the absence of standardized titles prevents direct mapping.

Long-range coreference chains. Individual artist biographies often span multiple paragraphs, resulting in coreference chains that traverse paragraph and even biography boundaries.

Effective resolution requires the system to persistently track and update entity representations over extensive contextual spans.

Deictic and generic expressions. Vasari routinely employs deictic and generic expressions such as 'the work', 'this figure', 'the said panel', or 'it' to refer either to specific entities or to artistic practice in general. For example, in statements that address the qualities of painting, sculpture, or architecture as processes, the referent is not a discrete artwork but rather the domain of craftsmanship itself. The system must distinguish between coreferential mentions that require resolution, potentially as implicit entities, and generic references that should be excluded from entity modeling. This distinction is exactly the contested boundary that the implicit-entity-recognition literature surveyed in Section 5.1.4 sets out to articulate, and ObliquER's four-way mention taxonomy is in part a response to the under-specification of that boundary in prior schemata.

Long-tail and out-of-knowledge-base entities. Many artworks described by Vasari are either absent from Wikidata or present with very sparse metadata, often not referenceable from the implicit mentions in the text. For example, the implicit artwork mention in this utterance "at the end of the sixteen months that Michelagnolo toiled over the work it was placed on the frontispiece in the façade of the Church of S. Petronio, as has been related; and we have also spoken of its size" is "the frontispiece in the facade of the Church of S. Petronio", which essentially refers to Wikidata entity `Q3769677`, a sculpture without a label definition. In EL, candidate generation methods that rely on label matching achieve low recall for these entities, whereas recall can be improved by constraining the candidate space using known artist metadata (e.g., the Wikidata `P170` creator property). This pattern is the CH instantiation of the long-tail and OOKB problem surveyed in Section 5.1.8, where canonical retrieval signals degrade precisely on the entities of most interest to the domain expert.

Language variations. The source text is a 20th-century English translation of a 16th-century Italian original. Artist names appear in Italianized, Latinized, and anglicized variants. Take, for example, "Michelagnolo Buonarroti", the form Gaston du C. de Vere uses for Michelangelo in his translation, whereas KBs typically refer to him as Michelangelo in English.

7.4.2 Relation to the Viewsari ontology

ObliquER does not operate in isolation. It is embedded within the Viewsari pipeline and is connected to both the Viewsari ontology (Section 7.3) and the requirements engineering process

that shaped it (Section 7.3.2.1). The formal task definition of Section 7.4.3 maps directly onto concepts in the current version of the Viewsari ontology.

Each mention m_j with type $\theta(m_j) \in \{\textit{Explicit}, \textit{Implicit}, \textit{Coreferent}, \textit{Generic}\}$ corresponds to a `viewsari:mention` instance – and its appropriate subclass – typed simultaneously as `oa:Annotation` and `prov:Entity`:

- $\theta(m_j) = \textit{Explicit} \Rightarrow$ `viewsari:explicit mention` ($\sqsubseteq$ `viewsari:mention`); for artwork entities, further refined to `viewsari:explicit artwork mention`.
- $\theta(m_j) = \textit{Implicit} \Rightarrow$ `viewsari:implicit mention`; for artworks, `viewsari:implicit artwork mention`.
- $\theta(m_j) = \textit{Coreferent} \Rightarrow$ `viewsari:coreferent`. Coreferent mentions are linked to their antecedent mention via `viewsari:refers to`. Coreferents are preserved in the KG for analysis but are excluded from ObliquER precision and recall computation.
- $\theta(m_j) = \textit{Generic} \Rightarrow$ `viewsari:generic mention`. As with coreferents, generic mentions are in the graph but excluded from evaluation.

The dual `oa:Annotation` / `prov:Entity` typing of `viewsari:mention` is a deliberate design decision documented in Section 7.3.2.2: every mention is simultaneously a scholarly annotation anchored in the source text and a provenance artifact generated by a documented computational process. Accordingly, each mention instance carries:

- `oa:hasTarget` $\rightarrow$ a `doco:TextChunk` instance, which in turn carries `oa:hasSelector` $\rightarrow$ a `oa:TextPositionSelector` with `oa:start` and `oa:end` character-offset values.
- `oa:hasBody` $\rightarrow$ `oa:TextualBody` $\rightarrow$ `rdf:value` the surface form string (e.g., *"a panel he executed for the high altar"*).
- `oa:hasSource` and `prov:used` both point to the containing `doco:Paragraph` expression, so that the mention is anchored to its source passage at both the annotation and provenance levels.
- `prov:wasGeneratedBy` $\rightarrow$ a `viewsari:named entity recognition` activity instance recording the extraction agent, the model identifier, and the prompt template version used.

The entity cluster C_e from the formal task definition corresponds to the `viewsari:extracted content` individual (typed simultaneously as `prov:Entity`) to which all mentions in the cluster resolve, as illustrated in Figure 7.23. For artworks, this individual is typed additionally as `viewsari:artwork` ($\sqsubseteq$ `viewsari:extracted content`). EL results populate `owl:sameAs`

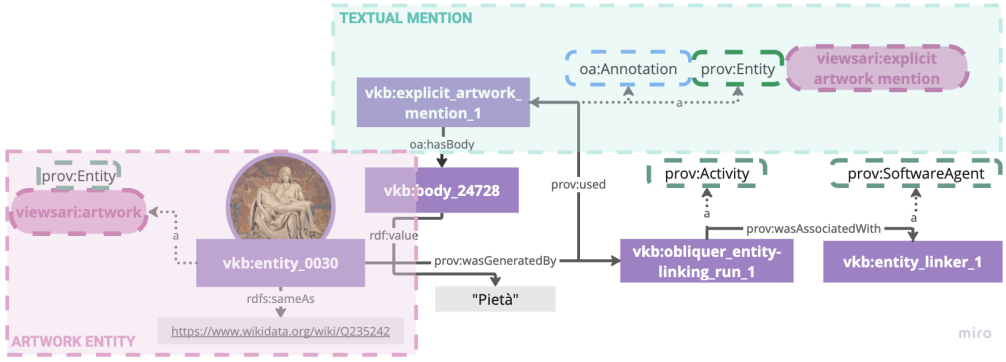


Figure 7.23: Ontology-level representation of the mention-to-referent mapping in Viewsari. A textual mention (`vkb:explicit_artwork_mention_1`, typed as `oa:Annotation`, `prov:Entity`, and `viewsari:explicit artwork mention`) is linked via `oa:hasBody` to a body node carrying the surface form "Pietà". The body resolves to the artwork individual `vkb:entity_0030` (typed as `viewsari:artwork` and `prov:Entity`), which is connected to its Wikidata counterpart via `owl:sameAs`. Linking provenance is captured by `prov:wasGeneratedBy`, pointing to the `vkb:obliquer_entity-linking_run_1` activity (typed as `prov:Activity`), which in turn is associated with the software agent `vkb:entity_linker_1` (typed as `prov:SoftwareAgent`) via `prov:wasAssociatedWith`. This separation makes recognition and linking provenance independently queryable at the ontology level.

between a Viewsari artwork individual and its Wikidata counterpart, or the artwork receives a Viewsari-internal link to become referenceable as an OOKB entity. The `prov:wasGeneratedBy` link from the artwork individual points to the corresponding `viewsari:entity linking` activity, distinguishing linking provenance from recognition provenance at the ontology level. In summary, the ontology defines the `viewsari:mention` hierarchy, the `viewsari:extracted content` superclass, and the typed activity subclasses introduced above. ObliquER populates these classes with instances derived from our case-study corpus (Section 7.5). The requirements engineering phase for Viewsari resulted in CQs that directly inform the design of the ObliquER framework, as they require representing artwork entities in the KG that cannot be derived solely from explicit canonical-title matching, such as `CQII.6` and `CQII.7`, quoted above.

In this context, ObliquER functions as a requirement-driven module whose scope and output schema are determined by the demands placed upon the ontology and KG, rather than serving as a generic extraction component. This entails, in particular, the necessity for transparency and replicability in the extraction process, as well as the systematic recording of provenance for each extracted mention, a requirement that motivated the annotation schema described in Section 7.4.6.

7.4.3 Formal task definition of entity recognition and linking

To establish a precise foundation for the recognition and linking tasks addressed by ObliquER, we present the formal task definition below. We begin by defining a document to be handled by the framework as d, an ordered sequence of paragraphs.

$d = \langle p_1, p_2, \ldots, p_n \rangle$.

Each paragraph p_i is, in turn, represented as a sequence of tokens.

A mention span $m = \langle p_i, s, e \rangle$ is defined as a contiguous subsequence of tokens within paragraph p_i, specified by a start offset s and an end offset e, both measured in character positions relative to the paragraph.

The corresponding text sub-string for a given mention span, referred to as the surface form $\sigma(m)$, serves as the basis for subsequent processing. The overall task is structured as two interdependent subtasks:

Entity recognition. Given a document d, the objective is to identify the complete set of entity mentions $\mathcal{M} = \{m_1, m_2, \ldots, m_k\}$ corresponding to a specified entity type τ; in this context, we focus on *Artwork*.

For each mention m_j in $\mathcal{M}$, the following assignments are required:

1. assign a *mention type* $\theta(m_j) \in \{$*Explicit*, *Implicit*, *Coreferent*, *Generic*$\}$;
2. assign an *entity identifier* $\varepsilon(m_j) \in \mathcal{E}$, where $\mathcal{E}$ is the

   set of unique real-world artworks referred to in d;
3. for coreferent mentions, record the antecedent identifier

   $\alpha(m_j) \in \mathcal{M}$, i.e. the earlier mention to which m_j co-refers.

A scalar per-mention confidence value $\kappa(m_j) \in \{$*high*, *low*$\}$ was envisaged in earlier drafts as a fourth output of this stage, intended to record the system's epistemic stance on each typed mention. The current pipeline does not emit such a value, and the ontology does not declare a corresponding data property; the epistemic-qualification role the value would have carried is instead borne by the mention typology $\theta(m_j)$ itself, which records the *manner* of identification rather than a scalar self-report. The deferral is documented as a limitation in Section 11.9.

Mentions that share the same entity identifier ε are grouped into an entity cluster, denoted as $C_e = \{m \in \mathcal{M} \mid \varepsilon(m) = e\}$.

Entity linking. Given the set of entities $\mathcal{E}$ and their associated clusters $\{C_e \mid e \in \mathcal{E}\}$, each entity e is to be linked either to a corresponding identifier in an external KB (in this case, Wikidata), or, if no suitable entry exists, assigned a Viewsari identifier to indicate it is OOKB. Formally, the linking function is defined as

$$\lambda : \mathcal{E} \longrightarrow \mathcal{K} \cup \{OOKB\}$$

where $\mathcal{K}$ denotes the set of Wikidata entity identifiers.

Candidate ranking. For each entity $e \in \mathcal{E}$, the retrieval stage returns a candidate set $K_e \subseteq \mathcal{K}$ of size at most k. The candidate-ranking step then selects either some $k_i \in K_e$ as the link target $\lambda(e)$ or returns *OOKB* when no candidate faithfully refers to e. ObliquER formulates this decision as a prompted task for an LLM based on the mention's surface form, the paragraph context, the candidate descriptors, and a role-based system instruction, and requires the LLM to produce either an index i or the literal *nil*. The ranking prompt is a few-shot prompt whose worked examples cover explicit, implicit, and OOKB scenarios.

Mention type definitions. We define the mention types used throughout ObliquER, formalized in the Viewsari ontology, as follows.

Explicit. The artwork is canonically identified by its surface form, a formal title, or a recognized traditional identifier (e.g., "La Gioconda", "The Battle of Marciano").

Implicit. A specific, recoverable artwork is described but not named.

Implicit mentions are, for example:

- *descriptive passages* that reference a commission, location, iconographic subject, or material without stating a canonical title (e.g., "a panel for the friars of Santa Croce, representing the Last Supper");
- *scene descriptions* identifying an artwork by its depicted content (e.g., "God dividing Light from Darkness" as a discrete scene in the Sistine Chapel);
- *deictic expressions* that are the primary introduction of an artwork in the text (e.g., "the statue", "the vault"), i.e., expressions that function as first mentions and therefore bear the referential weight of an implicit mention rather than a coreference.

The principal distinguishing criterion for implicit mentions is recoverability, in that the mention must refer to a unique, concrete artwork that can, in principle, be identified on the basis of contextual evidence.

Coreferent. The surface form refers to an artwork introduced earlier in the document, whether through an explicit title or an implicit description. Coreferent mentions do not introduce new entities; they extend existing coreference clusters. Typical coreferent forms in Vasari's text include pronouns ("it", "which"), definite noun phrases ("the painting", "the work"), and descriptive repetitions ("the same altarpiece").

Generic. A generic mention refers to an entity type in general, rather than to a specific individual. Generic mentions arise in artistic practice, for example, in the form of a craft process (e.g., 'works of painting' in the sense of the art of painting), rather than referring to a specific artwork.

To distinguish implicit first mentions from coreferent mentions, we rely on the presence or absence of an antecedent artwork mention within the local context window. If a deictic expression appears and no antecedent is available, either because the relevant artwork was introduced in a preceding paragraph outside the current context window or has not yet been introduced, we annotate the mention as `Implicit` rather than `Coreferent`.

Scope of coreference: biography-internal discourse threads only. ObliquER constrains coreference resolution to the biography as the fundamental discourse unit. A coreferential mention resumes a discourse thread initiated within the same biography; its antecedent must be identifiable within that biography, even if it falls outside the immediate context window and thus requires reconciliation at the entity-cluster unification stage rather than during local recognition (see Section 7.5). Coreference chains do not extend beyond the boundaries of unique biographies, even when two mentions in separate biographies refer to the same real-world artwork. Instead, the equivalence of this entity is determined by resolving it to the same entity from the EL step. We determine this restriction to be architectural and theoretical; ObliquER's sliding-window processing does not retain entity memory across biographies, so a mention in Botticelli's biography cannot access entities introduced in Ghirlandaio's biography. Theoretically, coreference presupposes a shared discourse context in which the antecedent remains accessible to the reader. This condition is satisfied within a single biography but not across biographies, which Vasari presents as self-contained and which readers may encounter in any order. Consequently, mentions of the same artwork in different biographies are annotated as independent explicit or implicit mentions, each with its own `entity_id`. Identity across biographical boundaries is established, as mentioned above, through shared identifiers; i.e., both mentions are linked to the same Wikidata item identifier (Q-number, QID). The KG population stage merges the corresponding entity nodes

accordingly (see Section 7.5). For example, Ghirlandaio's *St Jerome in His Study* is explicitly mentioned in his biography, while Botticelli's biography references the same painting as a pendant to his own *St Augustine*. In Botticelli's biography, *St Jerome* is annotated as a new explicit mention with its own `entity_id`, rather than being coreferential with the mention in Ghirlandaio's biography. Both mentions are linked to `wd:Q3463922`, and the KG records them as distinct annotation events that resolve to a single artwork node via `owl:sameAs`. This approach preserves the provenance of each mention, including its origin story, textual context, and extraction run, while ensuring the correct representation of entity identity in the KG.

This decision prioritizes the semantic content necessary for EL over strict linguistic coreference theory (Lata et al. 2022), reflecting our orientation toward the population of KGs rather than coreference evaluation alone.

7.4.4 Annotation strategy and ground truth construction

The annotation process is structured so that the resulting dataset serves as the GT, capturing the epistemic structure underlying Giorgio Vasari's *The Lives* and enabling evaluation of ObliquER's capability to extract and link artwork mentions, whether implicit or explicit.

The GT annotation corpus is constructed following requirements-based sampling principles as established in corpus annotation methodology (Pustejovsky and Stubbs 2012). Rather than going for an exhaustive annotation, the selection process is driven by explicit criteria, including representativeness across Vasari's periodization of the arts in the Italian Renaissance (Rubin Patricia Lee 1995); typological coverage of all four mention types (explicit, implicit, coreferent, generic); and the addition of negative examples to mitigate precision inflation during evaluation. This requirements-driven strategy reflects standard practice in annotation for low-resource or domain-specific corpora, where full-corpus annotation is infeasible and principled sampling constitutes the methodological norm. The resulting corpus of 232 annotated paragraphs, drawn from 16 biographies, functions as a controlled evaluation set, deliberately constructed to encompass the full spectrum of referential phenomena that the ObliquER pipeline is designed to address.

7.4.4.1 Scope of annotation: why artworks?

The ObliquER GT and extraction pipeline is currently scoped to artworks, understood as any discrete creative output described or alluded to in Vasari's text, including paintings, frescoes, sculptures, architectural works, and preparatory drawings. Essentially, the broader category of entity types relevant to *The Lives* is substantially wider than just artworks, spanning persons (artists, patrons, popes, merchants), locations (cities, churches, chapels, palaces), historical events

(commissions, deaths, artistic contests), and temporal markers. Several of these have been approached before by Santini et al. (2022) (cf. Section 5.2.4.2), and to integrate these results, the Viewsari ontology accommodates these classes as `viewsari:person` and `viewsari:place` alongside `viewsari:artwork`. The question is therefore not whether persons or locations matter to the KG, but why the GT construction and the ObliquER pipeline prioritize artworks as the primary annotation target. Three independent arguments converge on this choice.

Epistemological reasoning for the artwork scope. The central theoretical claim advanced in this dissertation is that KE from interpretive historical texts is fundamentally a process of construction rather than recognition (see Section 11.1). This claim is most evident when the distance between surface form and referent is greatest, that is, where identifying the entity in the text requires interpretive inference. For persons and locations in Vasari, this distance is often minimal, since named individuals are typically recoverable via label lookup in standard KBs, and locations are generally specified explicitly. As demonstrated by Santini et al. (2022, 2026), a classical BERT-based pipeline achieves satisfactory performance on these entity types precisely because they adhere to the surface-form-to-entity patterns on which such models are trained (thus, conforming to the PIEA). In contrast, for artworks, the gap is structurally significant. Vasari refers to artworks by their canonical titles when possible (or available at all). However, more frequently, he provides descriptions such as 'a figure of the Virgin with the infant Christ and several angels' rather than 'the Madonna of the Rocks', or 'a large panel for the high altar of the church' rather than 'the Portinari Altarpiece'. Identifying the referent in these cases requires the integration of contextual information, including location, commissioner, iconographic description, and period, which cannot be resolved through surface-form lookup alone. It is in these implicit and descriptive references to artworks that the epistemological challenge addressed by this dissertation is most acute, and it is precisely here that the novel contribution of ObliquER is situated. The decision to scope the GT to artworks is thus not a limitation of ambition but a methodological choice, focusing on the entity type for which the theoretical argument is empirically most rigorous.

In addition to the epistemological reasoning behind choosing artworks primarily, there is also an argument of feasibility. The process of manually annotating persons and locations at scale in *The Lives* is, in principle, possible but practically intractable within the resources of this dissertation project. The co-occurrence analysis in Section 6.4 identifies 1,061 distinct person names across the corpus; even the subset of 326 persons who appear in at least one co-occurrence pair poses a disambiguation challenge that would require art-historical domain expertise for every case of ambiguity. For artworks, by contrast, the annotation target is bounded by the sampled biographies (Section 7.4.4.4), and the primary reference resource, namely the Wagenbach critical edition (Nova 2004), which provides artwork identification notes keyed to Vasari's text, making the annotation task tractable for a trained annotator without requiring exhaustive Renaissance art

expertise. The Wagenbach edition functions not only as an annotation aid but as a GT anchor to harmonize the scholarly consensus on what Vasari is describing, including for works that lack Wikidata entries, enabling the OOKB cases to be populated reliably (Section 7.4.1).

Addressing the research questions. RQ2 addresses the recognition and linking of implicit and coreferential entity mentions within historical biographical prose. Although implicit references to persons are present in Vasari, including pronoun chains extending across multiple sentences or epithets such as "the divine" denoting Michelangelo, the structural challenge of implicit reference is most acute for artworks. While a person mentioned by nickname or epithet can often be resolved through knowledge-base search or coreference with a prior explicit mention in the same biography, an artwork referenced solely through a descriptive phrase, such as "a cartoon for a door-hanging that was to be executed in Flanders, woven in gold and silk, to be sent to the King of Portugal, of Adam and Eve sinning in the Earthly Paradise," lacks not only a canonical title but, accordingly, any corresponding KB entry. Its identification depends entirely on the surrounding descriptive context. This constitutes the OOKB problem in its most pronounced form and is predominant among artwork entities in the Vasari corpus. Consequently, the GT is defined with respect to the entity type for which RQ2 presents the greatest methodological difficulty, rather than the type for which it is most easily addressed.

However, we consider it worth stating explicitly that this scoping does not reduce the importance of other entity types, such as persons or locations. We represent a part of it through the co-occurrence layer. It does not imply that ObliquER cannot be applied to other entity types; we aim to demonstrate this in the transferability study on the FactGrid (Simons 2024) examples (cf. Section 10.1). What the scoping decision does imply is that the methodological and theoretical contribution of this dissertation, provenance-aware modeling of interpretive extraction, is demonstrated on an entity type and corpus combination where that contribution is most needed and very difficult to achieve.

7.4.4.2 Corpus construction and sampling strategy

As previously outlined, all knowledge-extraction steps in ObliquER are based on the 1912 English translation of *The Lives*, available in the Project Gutenberg corpus (Vasari 2008), which is distributed across ten volumes, each containing a subset of the biographies. The text was harvested paragraph by paragraph from the Gutenberg HTML representation using a dedicated scraper.[17] Each paragraph is stored along with its page number and paragraph index within the

[17] Content extraction scripts: `https://github.com/ISE-FIZKarlsruhe/viewsari/tree/dev/src/extract_content`.

page. This granularity corresponds to the printed book's paragraph structure and serves as the unit of processing for the ObliquER pipeline:

Listing 7.1: Storage format for harvested paragraphs.

```
page, paragraph, text
3,  1,     ``While the most noble and ...''
4,  1,     ``and arts -- namely, in painting, ...''
4,  2,     ``There was born a son, then, ...''
```

The GT provides few-shot examples, a gold-standard dataset for recognizing implicit entities in historical documents, and a basis for systematic evaluation.

Essentially, the biographies span from the pre-Renaissance proto-naturalists of the Trecento (e.g., Cimabue, Giotto) to the Mannerist painters of the mid-16th century (e.g., Pontormo). The referential conventions, the density of artwork descriptions, and the proportion of artworks with canonical titles all vary across this temporal range (cf. the uneven distribution of co-occurrences in different volumes in Section 7.2.2 and artworks in Figure 7.27). Our GT aims to sample a variety of perspectives that Vasari introduces, so taking excerpts from a single period would risk overfitting to period-specific linguistic patterns, for example, when targeting popular figures like Michelangelo, who are represented and referenced differently across their biographies (Nova 2004, Burioni 2017).

The GT is constructed to include examples spanning an exhaustive range of artworks mentioned in *The Lives*, including paintings, frescoes, sculptures, architectural projects, and decorative objects, ensuring that the framework is not limited to panel paintings, which are often foregrounded in conventional conceptions of artworks.

The sampling strategy follows Bogetić et al. (2022), who point out that choices about source, scope, segmentation, and annotation in DH corpora influence both the questions researchers can ask and the answers they can find. As a result, compiling the corpus is part of the research argument itself. The GT here is not a neutral sample. Instead, it is a structured selection meant to highlight the key phenomena this dissertation examines, namely how artwork references are constructed through implicit description, how OOKB entities are recognized, and how coreferential chains link explicit and implicit mentions across paragraphs.

In case-based research, sampling decisions of this kind operate so that the relevant question is not only whether the sample is representative of a population, but whether it allows the mechanisms of interest to be examined under the conditions that matter (Crowe et al. 2011). The unit of generalization is the case, not the paragraph. Paragraph selection within *The Lives* therefore

aims at analytic coverage of the phenomena that make the case informative, so the separation in the Renaissance eras Vasari describes, juggling variation in artistic medium, variation in artist prominence, and the distribution of mention types, rather than at proportional representation of paragraph frequencies in the source text. The prominence axis is based on the corpus-overview analysis. We found that the main high-frequency cluster identified by the co-occurrence rankings (Figure 7.3, Figure 7.4) is exactly what the stratified design limits, rather than repeats. This means the evaluation set focuses on less well-known figures rather than reiterating the most prominent names.

Sampling size of paragraphs from biographies. At first, we aimed for a sample size of 400 paragraphs. The 400-paragraph target was chosen as the smallest corpus size compatible with reliable per-mention-type evaluation at the observed density of around 6 mentions per paragraph; 400 paragraphs yield several hundred instances of every mention type, becoming comparable, on a magnitude level, to gold corpora of close approaches like ExtrART (Santini et al. 2022) and HIPE-2020/-2022 (Ehrmann et al. 2023). However, as Figure 7.25 and Figure 7.24 show, this size is not feasible for Vasari's work. Michelangelo's biography alone accounts for over 31% of paragraphs in the initial 400-paragraph target, and a GT weighted accordingly would produce evaluation metrics that reflect the linguistic patterns of a single exceptional text rather than the heterogeneity of Vasari's prose across the corpus. The argument for stratification follows the reasoning of Olieman et al. (2017) for corpus selection in DH EL, where the goal is not to reflect the frequency distribution of mentions but to surface the long tail and the structurally distinct cases that frequency-weighted sampling would obscure. They observe that frequency and relevance can be inversely correlated in DH corpora, where rare cases often carry more analytic weight than the common ones. Stratifying paragraph selection across biographies is therefore not a departure from representativeness but a recognition that representativeness in this context means coverage of the interpretive variety the case is meant to illustrate.

Returning to the selection of paragraphs, the skew arises directly from the structure of *The Lives* itself, since Vasari devoted far longer biographies to artists he considered the apex of the Renaissance progression, meaning that a proportional sampling strategy amplifies this canonical hierarchy (see Figure 7.24). To address this, the sampling strategy was revised to target a balanced distribution of 10–20 annotated paragraphs per biography, with a soft cap that prevents any single biography from exceeding roughly 11% of the total corpus. For several shorter biographies, namely Gherardo (5 paragraphs available), Cosimo Rosselli (4), and Pietro Cavallini (7), this range is already exhaustive of the available sampled material, and all paragraphs are included. For the longer biographies, paragraphs were selected to represent the full diachronic and typological range of the biography: early, middle, and late sections; varied artwork types (panel paintings, frescoes, sculptures, architectural works); and a mix of explicit, implicit, and coreferential mention

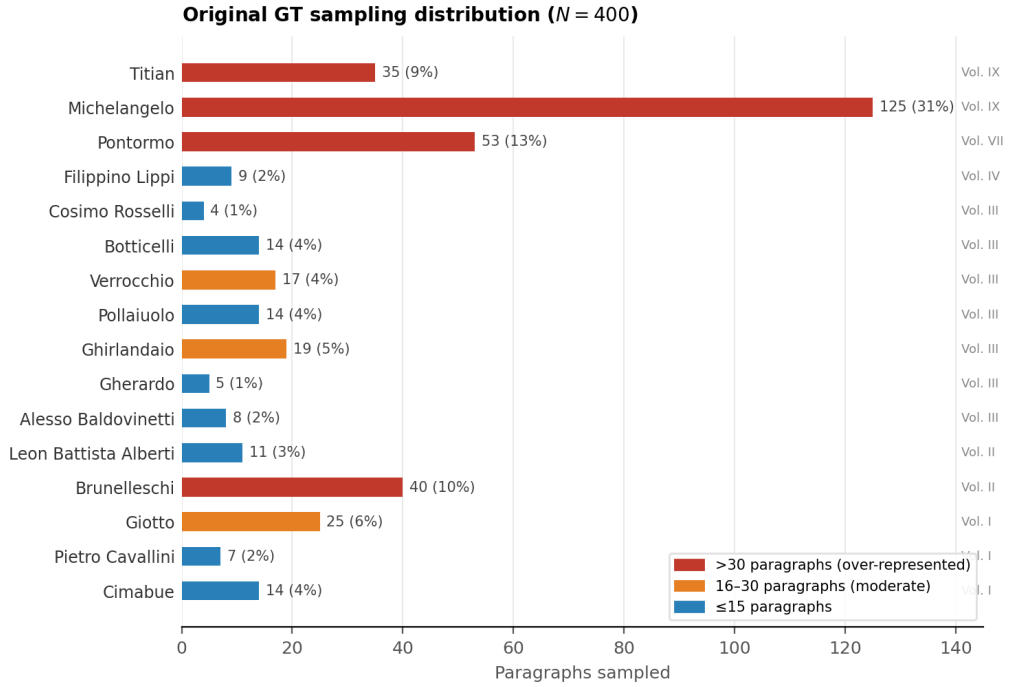


Figure 7.24: Original GT sampling distribution ($N = 400$). Biographies are ordered chronologically by volume. Color encodes over-representation: red bars (> 30 paragraphs) indicate biographies whose share of the total corpus exceeds the per-biography cap adopted in the revised strategy.

types. Michelangelo's biography is treated as an anchor case, retaining its full prior annotation run of 49 paragraphs to reflect its role as the primary schema-development and test biography. The stratified design targets a balanced evaluation set of 221 paragraphs capped at 11% per biography (see Figure 7.25, with the Michelangelo biography as an exception); the realized corpus comprises 232 annotated paragraphs across 16 biographies (Table 7.5). Of these, the six demonstration and modeling-example paragraphs are held out from the recognition-evaluation partition (Section 7.4.7), so the recognition results of Chapter 9 are computed on the remaining 980 of the 1,010 implicit-and-explicit mentions. For the robustness of the GT, the inclusion of paragraphs containing no artwork mentions – referred to as *negative examples* – is a design requirement rather than an incidental feature. Crowe et al. (2011) emphasize that case-internal evidence selection should anticipate what would count as a boundary case, not only what would confirm the pattern under study. Paragraphs that contain phrases superficially resembling artwork references or no such cases at all function as such for the artwork annotation schema. Their inclusion prevents the GT from being constructed solely from positive instances, which would

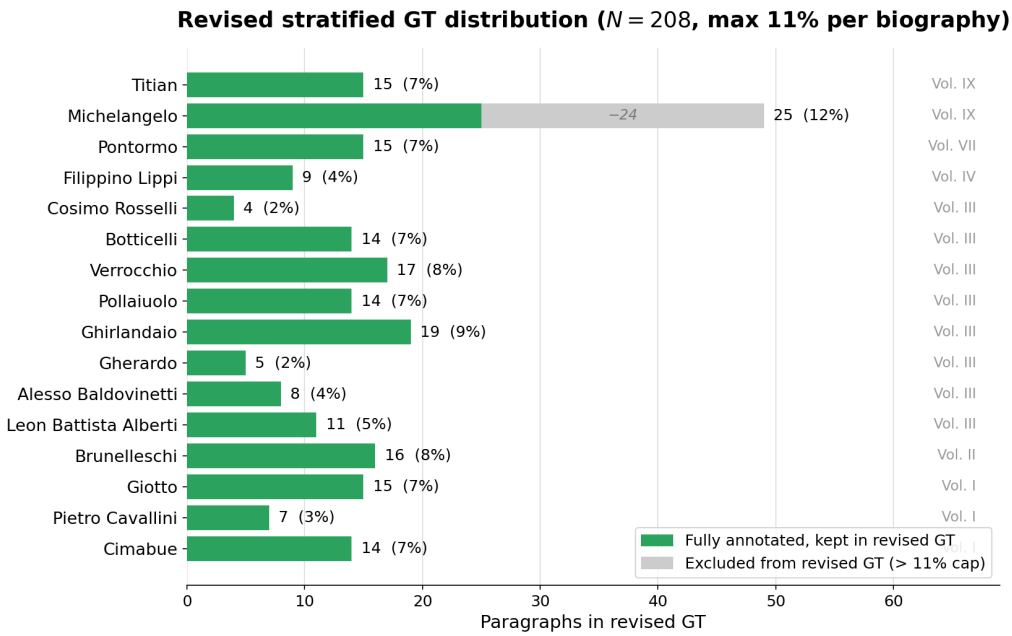


Figure 7.25: Revised stratified GT distribution ($N = 208$, maximum 11% per biography, Michelangelo biography is an exception). All sixteen biographies are now fully annotated and retained in full (green bars). The only exception is Michelangelo, the anchor biography, for which 49 paragraphs were annotated during the schema-development phase; it is capped at 25 paragraphs (≈12%) in the revised GT, with the gray extension indicating the 24 excess paragraphs excluded from the revised GT and reserved for the development and ablation sets.

render the schema's specificity untestable.[18] Additionally, our sampling strategy aims to achieve a balanced distribution across *Explicit*, *Implicit*, and *Coreferent* mentions. Because the absolute frequency of each mention type varies considerably across biographies, balancing is achieved at the corpus level rather than within individual biographies.

7.4.4.3 Biography selection criteria

Biographies are selected to represent each of the three main volumes of *The Lives*, which Vasari himself organized to reflect a narrative of the 'progressive perfection of the arts'. The periodization of *The Lives* has been extensively discussed in the secondary literature and is treated

[18] Volume numbering follows de Vere's ten-volume Gutenberg edition. Gaps in the sequence (Volumes V, VI, VIII, and X are unrepresented) reflect the decision to sample biographies according to Vasari's periodization of the three eras rather than de Vere's later editorial division into ten volumes: `https://github.com/ISE-FIZKarlsruhe/ObliquER/tree/main/data/viewsari/ground_truth`

as foundational in standard art-historical reference works (Baxandall 1971, Murray and Murray 1963, Cole 1983, Blunt 1962).

- **Part I (the First Age** covers artists from Cimabue and Giotto, whom Vasari regards as the first restorers of naturalistic representation after the Byzantine period. Scholarship has treated this part as documenting a formative, pre-canonical stage in which biographical information is often sparse and artwork attributions tentative (Rubin Patricia Lee 1995, Baxandall 1971).
- **Part II (the Second Age** addresses masters like Brunelleschi, Donatello, Ghirlandaio, Botticelli, and their contemporaries, whom Vasari credits with the systematic conquest of perspective, anatomy, and design. The biographies become substantially longer and the network of artists denser (Murray and Murray 1963, Cole 1983).
- **Part III (the Third Age** culminates in Leonardo, Raphael, and Michelangelo, the last of whom Vasari explicitly crowns as the summit of human artistic achievement. It extends into Mannerism with figures such as Pontormo (Cole 1983).

However, temporal coverage alone does not suffice as a selection criterion. Additionally, we accounted for artist prominence, since the length of biographies and their density align with Vasari's view of prominence. In contrast, shorter biographies of less prominent artists yield a different distribution of mention types and a higher proportion of OOKB entities. Both ends of this spectrum are represented in the selection.

Based on these criteria, we select the biographies shown in Table 7.5 for inclusion in the Viewsari GT.

7.4.4.4 Sampling procedure and baseline corpus for the ground truth annotation

Paragraph selection for the GT corpus follows a stratified, reproducible procedure that operates on the per-volume paragraph tables produced by the preprocessing pipeline of Section 7.4.8.1. For each selected biography, paragraphs falling within the biography's page range – and above a minimum length threshold that suppresses short transitional fragments – are partitioned into two pools. *Candidate positives* are paragraphs containing at least one term from a controlled vocabulary of artwork-related signals (e.g., *painted*, *fresco*, *altarpiece*, *sculpture*, *chapel*, *design*); *candidate negatives* are their complement. The partition is deliberately heuristic. A candidate negative may contain artwork mentions that the keyword filter does not detect, and a candidate positive may describe artistic activity without referring to a discrete artwork entity. The gold-standard label is therefore assigned exclusively by the human annotator in INCEpTION, and

Table 7.5: Biographies selected for the Viewsari GT corpus, ordered chronologically by artistic period and volume. *Rev.* gives the number of paragraphs in the GT corpus; *Neg* counts heuristic negative-example candidates; *Avg W* reports mean paragraph word count. Michelangelo Buonarroti (*Rev.* = 49) reflects the full prior annotation run rather than the 10–20 stratified cap applied to the remaining biographies (*).

Artist	Vol.	Period	Orig	Rev.	Neg	Avg W
Giotto di Bondone	I	Proto-Renaissance (Trecento)	34	15	2	372
Giovanni Cimabue	I	Proto-Renaissance (Trecento)	14	14	3	222
Pietro Cavallini	I	Proto-Renaissance (Trecento)	7	7	0	228
Filippo Brunelleschi	II	Early Renaissance	43	16	1	505
Alesso Baldovinetti	III	Early Renaissance	8	8	1	195
Andrea del Verrocchio	III	Early Renaissance	17	17	0	234
Antonio & Piero Pollaiuolo	III	Early Renaissance	14	14	2	194
Cosimo Rosselli	III	Early Renaissance	4	4	0	320
Domenico Ghirlandaio	III	Early Renaissance	19	19	0	315
Gherardo di Giovanni	III	Early Renaissance	5	5	0	199
Leon Battista Alberti	III	Early Renaissance	11	11	1	216
Sandro Botticelli	III	Early Renaissance	14	14	4	220
Filippino Lippi	IV	High Renaissance	9	9	1	320
Jacopo da Pontormo	VII	Mannerism	64	15	1	261
Michelangelo Buonarroti*	IX	High Ren. / Mannerism	125	49	5	342
Tiziano Vecellio (Titian)	IX	High Ren. / Mannerism	35	15	1	287
Total				**232**	**22**	

the heuristic flag is preserved alongside the final label so that the sampling provenance remains traceable from the raw corpus to the evaluation set.

The vocabulary of positive signals was constructed from the subclasses of Wikidata's *work of art* (Q838948), supplemented by artwork annotations recorded in the German Wagenbach edition (Section 7.4.4.5).[19] A paragraph is treated as a candidate negative if none of these signals appear in its lowercased text. The filter is conservative by design: it admits false positives rather than suppressing rare or implicit references to artworks.

Within each biography's filtered pool, a two-pool draw approximates a target negative fraction of 30%. The negative quota is computed against the per-biography paragraph target and capped by pool availability; the remainder is filled from the positive pool. Where the positive pool

[19] SPARQL query: `?subclass wdt:P279 wd:Q838948`.

is itself smaller than expected, the negative quota is reduced in turn rather than inflating the total. This mechanism explains the 9.5% observed negative fraction in Table 7.5 despite the 30% target. Several biographies have too few positive paragraphs to meet the planned quota, and the cap absorbs the shortfall to preserve the per-biography count. The procedure is reproducible end-to-end. A fixed random seed is propagated to all draws, biography names are normalized and matched against the page-range index before sampling, and the resulting selection is sorted into reading order before being passed to the annotation environment (Section 7.4.5).

Revised stratified design. As discussed above (Section 7.4.4.2), the initial aggregate target of 400 paragraphs was subsequently revised to a per-biography cap of 10–20 paragraphs, with Michelangelo retained at 49 as the anchor biography. The revision is layered as a post hoc selection over the original 400-paragraph sample, with each paragraph carrying its membership in the evaluation corpus, along with an annotation status indicating whether it has been annotated, is in progress, has been reserved for prompt development and ablation experiments, or falls outside the revised selection entirely. This layered design preserves the full original sample as a reusable resource while explicitly distinguishing the evaluation corpus from the ablation set.

Corpus report. Table 7.5 shows the composition of the revised GT corpus with all 16 biographies, arranged by artistic period and volume. For each biography, it provides the original paragraph count from the 400-paragraph pool (Orig), the number kept in the revised stratified corpus (Rev.), the number of heuristic negative-example candidates (Neg), and the average paragraph word count (Avg W). Table 7.6 shows summary statistics for both the volume and the entire corpus. Volume III comprises the largest part of the corpus (39.7%), as most biographies sampled from the Early Renaissance are included here. This period has the highest density and variety of artworks in *The Lives*. The overall heuristic negative rate for the corpus is about 9.5%, which is lower than the original 30% target. This is mainly because many small biographies, when kept in full, have few or no paragraphs that meet the negative filter, and the per-biography cap tends to remove more paragraphs from larger biographies, which were better at reaching the negative target. On average, paragraphs in the corpus are 295 words long, which reflects the annotation effort needed for each unit. Because implicit entity recognition is complex, a paragraph of this length usually has between zero and four artwork mentions that can be annotated, though this varies a lot between periods.

Table 7.6: Corpus-level statistics for the Viewsari GT as realized ($N = 232$ paragraphs, 16 biographies). The negative fraction is computed over the GT; the heuristic candidate-negative flag marks paragraphs likely to contain no artwork mention (see Section 6.4.1). Volume distribution reflects Vasari's own three-part narrative of the progressive perfection of the arts.

Volume	Contents (period)	Paragraphs	Neg.	Share
Vol. I	Proto-Renaissance (Trecento)	36	5	15.5%
Vol. II	Early Renaissance (architecture)	16	1	6.9%
Vol. III	Early Renaissance (painting / sculpture)	92	8	39.7%
Vol. IV	High Renaissance	9	1	3.9%
Vol. VII	Mannerism	15	1	6.5%
Vol. IX	High Renaissance / Mannerism	64	6	27.6%
Total		**232**	**22**	**100%**
Additional corpus statistics				
Negative examples (heuristic)		22		9.5%
Mean paragraph length (all)	words	295		
Mean paragraph length (Trecento)	words	286		
Mean paragraph length (Early Ren.)	words	279		
Mean paragraph length (High Ren./Mannerism)	words	317		
Biographies		16		
Largest single biography share (Michelangelo)		49		21.1%

7.4.4.5 The Wagenbach critical edition as primary annotation reference

All manual annotation decisions in the Viewsari GT were made against the de Vere English translation of *The Lives* as the primary source text. However, de Vere's 1912 rendering, while authoritative for the English digital edition used as the ObliquER corpus, does not provide the scholarly apparatus – critical footnotes, iconographic commentary, artwork identifications, and bibliographic cross-references – that is indispensable for resolving the disambiguations described in Section 7.4.6.4. The principal secondary reference used throughout the annotation process is the *Edition Giorgio Vasari* published by Verlag Klaus Wagenbach, Berlin (Nova 2004). This series presents individual biographies in German translations, accompanied by an extensive critical apparatus in which the authors provide detailed annotations to the text. The series is particularly valuable for the present study because the editorial annotations frequently identify artworks mentioned by Vasari with their surviving counterparts, also resolving ambiguities through reference

to primary documents and archival sources, and flag cases where Vasari's account is considered unreliable or contested by modern scholarship.

The following nine volumes of the *Edition Giorgio Vasari* were consulted directly during annotation:

- *Das Leben des Cimabue, des Giotto und des Pietro Cavallini* (Vasari 2015)
- *Das Leben des Domenico Ghirlandaio und des Gherardo di Giovanni* (Vasari 2014)
- *Das Leben des Verrocchio und der Gebrüder Pollaiuolo* (Vasari 2012a)
- *Das Leben des Brunelleschi und des Alberti* (Vasari 2012b)
- *Das Leben des Filippo Lippi, des Pesello [. . .] und Domenico di Michelino* (Vasari 2011)
- *Das Leben des Sandro Botticelli, Filippino Lippi, Cosimo Rosselli und Alesso Baldovinetti* (Vasari 2010)
- *Das Leben des Michelangelo* (Vasari 2009)
- *Das Leben des Tizian* (Vasari 2005)
- *Das Leben des Pontormo* (Vasari 2004)

We consulted the critical apparatus when an artwork description in de Vere was insufficiently specific to enable a confident entity identification from the text alone – particularly for implicit artwork mentions. To some extent, the Wagenbach footnotes provided contextual grounding for these cases; however, not all artworks are described in the footnotes. Moreover, when the Wikidata search (anchored to the dump of mid 2025; see Section 7.4.5) returned multiple candidate items, we used the footnotes as 'ranker'. In said cases where the Wagenbach editions acknowledge uncertainty or missing information, we mark this mention as a reference to an OOKB entity (see the Titian example in Section 7.4.6.4).

7.4.5 Annotation environment

We annotate the paragraphs primarily in INCEpTION (Klie et al. 2018); however, we first annotated a set of mentions using LabelStudio but discarded that approach due to LabelStudio's limited linking interface.[20] The annotation for the Michelangelo biography was performed primarily by the author and Cedric Möller (co-author of the ObliquER work under review) and

[20] LabelStudio: `https://labelstud.io/`.

cross-referenced with the German critical-historical Wagenbach edition (Vasari 2009) to resolve ambiguous references.

Wikidata as the external knowledge base for manual annotation. All EL decisions made during manual annotation rely on Wikidata as the sole external KB (Vrandečić and Krötzsch 2014). When an annotated artwork entity can be matched to a Wikidata item, the corresponding identifier is recorded in the `wikidata_id` field of the EL annotation record. Entities for which no matching Wikidata item could be found at annotation time are recorded as OOKB and assigned a Viewsari-internal IRI via the INCEpTION OOKB mechanism (see Section 7.4.6).

Surface form selection criteria. We select surface forms for annotation based on clear criteria that define span boundaries and help decide if a phrase should be marked as an artwork mention. For example, paragraph 58 of the Michelangelo biography in the de Vere translation (Vasari 2008)[21] discusses the commission of the Last Judgment, Michelangelo's work on the tomb of Julius II at the same time, and an earlier devotional fresco in S. Gregorio based on a related design. We chose this paragraph because it contains all four mention types listed in Table 7.8. The full list of annotated mentions for this paragraph appears in Table 7.7, and the paragraph text with each annotated surface form in bold is shown below.

> Now, just as Michelagnolo was about to have **the statues** carried into execution, at that very time the Pope took it into his head to have him near his person, being desirous to have **the walls of the Chapel of Sixtus** painted, where Michelagnolo had painted **the vaulting for Julius II**, his nephew. On the principal wall, where the altar is, Clement wished him to paint **the Universal Judgment**, to the end that he might display in that scene all that the art of design could achieve [. . .]; of which inventions it was found that Michelagnolo many years before had made **various sketches and designs**, **one of which was afterwards carried into execution in the Church of the Trinità at Rome** by a Sicilian painter, who stayed many months with Michelagnolo, to serve him and to grind his colours. **This work**, painted in fresco, is in the Chapel of S. Gregorio, in the cross of the church [. . .]. While Michelagnolo was directing the preparation of **the designs and cartoons of the Last Judgment on the first wall**, he never ceased for a single day to be at strife with the agents of the Duke of Urbino, by whom he was accused of having received sixteen thousand crowns from Julius II for **the tomb**. This accusation was more than he could bear, and he desired

[21] Volume 9, file `michelangelo_enriched.json` in the Viewsari GT at `data/viewsari/ground_truth/9/`.

to finish **the work** some day, although he was already an old man, and he would have willingly stayed in Rome to finish **it** [. . .]. (Vasari 2008)

Table 7.7: Entity inventory for paragraph 58 of the Michelangelo biography (Volume 9, `michelangelo_enriched.json`, 11 mentions). All four mention types defined in Table 7.8 appear. Two coreferents (the statues, the tomb) point to antecedents introduced in earlier paragraphs and are resolved against the cross-paragraph cluster in the post-processing step of Section 7.4.8.1. The generic mention of various sketches and designs is directly contrasted with the immediately following implicit mention, which singles out a single recoverable artwork from the generic set. Three surface forms (the designs and cartoons of the Last Judgment on the first wall, the work, it) anchor the coreference chain of `artwork_56`, demonstrating the implicit → coreferent transition that is the most frequent recognition pattern in the corpus.

`entity_id`	**Mention type**	**Surface form**	**Identity / KB anchor**
(cross-paragraph)	*Coreferent*	the statues	antecedent in paragraph 57 (`m_0225`)
`artwork_52`	*Implicit*	the walls of the Chapel of Sixtus	Sistine Chapel walls – `Q123254487`
`artwork_53`	*Implicit*	the vaulting for Julius II	Sistine Chapel ceiling – OOKB
`artwork_54`	*Explicit*	the Universal Judgment	The Last Judgment – `Q567861`
(none)	*Generic*	various sketches and designs	generic collection; excluded from core eval.
`artwork_55`	*Implicit*	one of which was afterwards carried into execution in the Church of the Trinità at Rome	painting in Chapel of S. Gregorio – out-of-knowledge-base
`artwork_55`	*Coreferent*	This work	coref to `m_0233` (the S. Gregorio fresco)
`artwork_56`	*Implicit*	the designs and cartoons of the Last Judgment on the first wall	preparatory cartoons – OOKB
`artwork_1`	*Coreferent*	the tomb	cross-paragraph: tomb of Julius II (`m_0012`)
`artwork_56`	*Coreferent*	the work	coref to the Last Judgment cartoons (`m_0235`)
`artwork_56`	*Coreferent*	it	coref to the Last Judgment cartoons (`m_0237`)

For *Explicit* mentions, the full title or the complete noun phrase naming the artwork is selected, including any embedded descriptors (e.g., *The Entombment of Christ* rather than *Entombment*). For *Implicit* mentions, the minimal but complete descriptive phrase that uniquely identifies the referent within its context is selected (e.g., *a panel for the friars of Santa Croce, representing the Last Supper*). Descriptors that are redundant given the broader context may be omitted from the span, but the span must always include sufficient material to reconstruct the referent in isolation.

For *Coreferent* mentions, the minimal referring expression is selected (e.g., *it*, *the painting*, *this work*).

Sometimes, Vasari's prose includes one artwork reference inside another, such as a title within a longer description. In these cases, both the inner title and the outer phrase are marked as separate mentions that point to the same entity. If overlapping spans refer to different entities, they are annotated separately, and their entity ID fields show their different referents. Phrases that refer to a group of artworks, like 'many paintings in Florence' or 'his early works,' and not to a single, specific artwork, are thus labeled as *Generic Mention*, but this type is not included in the main evaluation for precision and recall.

References to architectural works, mostly buildings, are treated differently. We annotate a building as an artwork mention if it is a recoverable, named architectural work, regardless of whether Vasari describes it for its aesthetics, location, or function. This also applies when a building is only mentioned as a place for another artwork, like 'placed at the head of the altar of S. Jacopo in Pistoia,' or as a functional reference, such as 'in the church of Santa Croce,' as long as the building is a known, identifiable architectural or composite artwork. However, if the reference is too generic and cannot be linked to a specific named work, like 'in a chapel' or 'in the refectory,' and there is no further context, we do not annotate it as an implicit entity. Instead, we tag it as a 'generic mention'. This adheres to the comments in the Wagenbach edition, where buildings are also treated, even if only mentioned by Vasari, since references to them establish their relevance to the Italian Renaissance. More on this follows in the subsequent sections.

7.4.6 Annotation guidelines

We designed these guidelines to formalize the creation and extension of the Viewsari GT. They specify the annotation schema, rules for span selection and type assignment, and handling of edge cases. The guidelines are versioned and maintained with the GT data in the project repository.[22] They primarily serve as a reference for human annotators and also inform the prompt design discussed in Section 7.4.8. These guidelines explain how to annotate mentions of artworks, such as paintings, frescoes, altarpieces, sculptures, buildings, and decorative objects, in Vasari's *The Lives*. Because many artworks lack standard titles, we pay special attention to implicit mentions, in which artworks are described rather than named.

This version focuses only on the Artwork entity type. Mentions of other types, such as Artist, Patron, Place, and Material, will be included in future updates to the guidelines.

[22] Viewsari and ObliquER GT repository files: `https://github.com/ISE-FIZKarlsruhe/ObliquER/tree/main/data/viewsari/ground_truth`.

7.4.6.1 Mention types

As mentioned before, we differentiate between the following types of artwork mentions:

Table 7.8: Mention types in the ObliquER annotation schema.

Type	Definition	Example
Explicit	The artwork is named or clearly identified by a title (including conventional nicknames or traditional identifiers).	"He painted *The Battle of Marciano"*.
Implicit	A specific artwork is described but not named; the referent is recoverable from context, commissioner, location, or iconographic description.	"He painted a panel for the friars of Santa Croce."
Coreferent	Refers to a previously mentioned artwork (same entity), using a pronoun, definite noun phrase, or partial repetition.	"It was placed in their refectory."
Generic	Refers to an undifferentiated group or unspecified category of artworks rather than a single recoverable referent; no `entity_id` is assigned.	"some sculptures", "his works", "certain other works".

7.4.6.2 Annotation fields

Each annotation record must contain the fields listed in Table 7.9.

Knowledge base assignment during annotation. The `wikidata_id` field is populated using the INCEpTION built-in KB lookup, which was configured to query a local mirror of Wikidata during the annotation process. Annotators searched for candidate items by artwork title, creator name, and location, using the INCEpTION KB search interface. Where a match was found, the QID was recorded directly; where no match existed, the entity was flagged as OOKB and assigned a Viewsari-internal IRI. For ambiguous cases – where multiple Wikidata items were plausible candidates – the Wagenbach critical edition (Nova 2004) and specialist art-historical reference works were consulted as secondary evidence before a QID was committed (see Section 7.4.6.4 for a worked example from the Titian biography).

Table 7.9: Annotation fields for recognition records of artwork mentions.

Field	Description
`paragraph id`	Integer identifier of the source paragraph.
`mention id`	Unique string identifier for the mention (e.g., `m_0014`).
`type`	One of {`explicit artwork mention`, `implicit artwork mention`, `coreferent`, `generic collection mention`}.
`surface form`	The exact text span as it appears in the paragraph.
`entity id`	Identifier for the real-world artwork entity this mention refers to (shared by all mentions in the same cluster).
`refers to`	(For `coreferent` only) The `mention id` or surface form of the antecedent. `null` otherwise.
`start offset`	Character offset of the first character of the surface form within the paragraph.
`end offset`	Character offset of the character *after* the last character of the surface form.

7.4.6.3 Annotation instructions

Span selection. Select the surface form according to the span-boundary rules in Section 7.4.5. The surface form must exactly match the character sequence in the source paragraph without normalization or correction.

Type assignment. Assign the most specific applicable type. If a mention includes a canonical title (*Explicit*) within a longer descriptive phrase, annotate the inner explicit mention and the outer descriptive phrase separately as *Explicit* and *Implicit*, linking both to the same `entity id`.

Entity clustering. All mentions, explicit, implicit, and coreferent, that refer to the same real-world artwork must share the same `entity id`. Assign a new `entity id` when a new, distinct artwork is introduced.

Negative paragraphs. Paragraphs containing no artwork mentions must be included in the GT with an explicit `null` record.

7.4.6.4 Disambiguation and inference

To determine if an artwork mention is implicit, apply the following decision procedure:

1. **Unique, concrete referent?** If the phrase refers to a specific individual artwork that is recoverable from context, annotate as `implicit artwork mention`.
2. **Generic or collective?** If the phrase refers to a set of artworks or to an artist's output in general, do *not* annotate as `implicit artwork mention`. Optionally annotate as `generic mention`.

7.4.6.5 Coreference

Table 7.10 lists typical surface forms of coreferent mentions:

Table 7.10: Common coreference forms in Vasari's prose.

Form	Example	Comment
Pronouns	"It was placed in the refectory."	Most common; short-range reference to the most recently introduced artwork.
Definite noun phrases	"The painting was finished in 1554."	Refers to the previously introduced artwork.
Descriptive repeats	"The same work was later copied."	Combines a determiner (*the same*) with a generic head noun (*work*, *panel*, etc.).
Elliptical mentions	"This too he painted for the friars."	Deictic use of *this* or *that*; resolve by proximity and context.

Annotators must:

- Annotate every surface expression that refers back to a known artwork.
- Record the `refers to` field as the `mention id` (or surface form, if the `mention id` is not yet assigned) of the antecedent.
- Not create a new `entity id` for a coreferent mention. Select the most locally or contextually salient one.

7.4.6.6 Special cases

Sistine Chapel ceiling. The Sistine Chapel ceiling is discussed across many consecutive paragraphs in the Michelangelo biography. Individual scenes are annotated as separate *Explicit* or *Implicit* entities (e.g., *the scene of God dividing Light from Darkness*, *the Creation of the Sun and Moon*), while coreferent expressions referring to the ceiling as a whole (e.g., *the vaulting*, *it*) are linked to the ceiling entity `artwork_2` in the GT.

Architectural works. Buildings are marked as artwork mentions when Vasari presents them as creative achievements. The label is the most informative form in the text – a proper name (*the cupola of Santa Maria del Fiore*), a functional label (*the tribune of the Annunziata*), or a descriptive phrase (*the new sacristy he built for the Medici*). Vasari treats the design of buildings, fountains, conduits, and urban ensembles as creative achievements equal to paintings and sculptures, with the same iconographic and attributional detail. The annotation schema therefore treats any architectural work – church façade, dome, loggia, palace, or civic infrastructure such as the restored Acqua Vergine and its fountain on the Piazza de' Trevi – as a `viewsari:artwork`, provided Vasari refers to it as a distinct product of a known creator's design. The criterion is not the art-historical category but how the work is referenced. It counts as an artwork if it denotes a specific, individual work that could in principle be linked to a knowledge-base entity. Excluding architecture would impose a modern distinction the source rejects – the very title, *Le vite de' più eccellenti pittori, scultori e architettori*, places architecture on a par with painting and sculpture.

Designs and models. Cartoons, wax models, and preparatory drawings are annotated as *Implicit* mentions when Vasari describes them as specific objects with recoverable identities, even when they are described in terms of their relationship to a finished work.

7.4.7 Ground truth annotation process example in INCEpTION

This example comes from consecutive paragraphs in the Ghirlandaio biography (de Vere translation, Volume III, pp. 219–221; paragraph identifiers 248–250 in the merged corpus) as shown in the INCEpTION annotation interface. Figure 7.26 displays the annotation state of these paragraphs. Each colored span label matches one of the roles listed in Table 7.8. This passage was selected because it includes all four types, namely Explicit Artwork Mention, Implicit Artwork Mention, Coreferent, and Generic Mention. It also highlights several challenges in Vasari's prose, including dense references, overlapping mention spans, coreference chains, and works described

by function or location rather than by title. We choose Paragraph 248 (cf. Section 7.4.7) as an illustrative example for details on decisions made to annotate mentions as artworks.



Figure 7.26: INCEpTION annotation view for paragraphs 248–250 of the Ghirlandaio biography (de Vere Vol. III, pp. 219–221). Colored span labels indicate mention role: blue = Explicit Artwork Mention (with Wikidata QID where available), green/orange = Implicit Artwork Mention, yellow = Coreferent, purple = Generic Mention. Arcs labeled *Corefent Links* connect coreferent mentions to their antecedents. The paragraph header (`# paragraph_id | pages | candidate_negative | text`) is generated automatically from the merged-corpus CSV.

Paragraph 248 (p. 219). This introductory paragraph about Ghirlandaio's father, Tommaso, explicitly and implicitly mentions various artworks, despite being primarily biographical. Among those, examples are:

- **Implicit and Explicit Artwork Mention**: *the greater part of the silver votive offerings that were formerly preserved in the press of the Nunziata.* The span is annotated as both implicit and explicit because Vasari identifies a specific, recoverable collection of objects by their institutional location, which is itself an architectural building (the Santissima Annunziata, Florence); the Wikidata identifier Q2287001 is assigned to the church.

- **Implicit Artwork Mention and its Coreferent**: *the silver lamps of the chapel* and the pronoun *which*, linked back to the mention of the silver lamps of the chapel, via a *Corefent Links* arc.
- **Implicit Artwork Mention and another Coreferent**: *those ornaments worn on the head by the girls of Florence*, referred to subsequently by the coreferent *which* (arc visible in the interface). Vasari describes these as the garlands (*ghirlande*) that gave the family its nickname; they are a specific class of craft object recoverable from context, hence Implicit rather than Generic.
- **Generic Mention**: *an infinite number of them* – a collective, indefinite reference to the garlands as a category, not to a specific recoverable object.

Annotation observations from the screenshot. Several properties of the INCEpTION interface visible in Figure 7.26 are worth noting for methodological transparency.

As such, where multiple mentions overlap (e.g. the Nunziata), INCEpTION renders each label on a separate tier above the text. This stacking is a consequence of the annotation schema permitting both inner and outer spans on the same surface form. Additionally, coreferents lead to a particularly dense arc structure. The *Corefent Links* arcs visualize these chains and allow a verification of their completeness before export.

Disambiguation using secondary sources: a Titian example. A recurrent disambiguation challenge in the Titian biography arises from Vasari's tendency to name works collectively before describing individual canvases in detail. The following passage, drawn from the Titian biography (de Vere translation, Volume IX), illustrates a case in which the initial collective reference cannot be resolved without consulting sources external to the de Vere text:

> In the year 1514 Duke Alfonso of Ferrara had caused a little chamber to be decorated, and had commissioned Dosso, the painter of Ferrara, to execute in certain compartments **stories of Æneas, Mars, and Venus**, and in a grotto **Vulcan with two smiths at the forge**; and he desired that there should also be there pictures by the hand of Gian Bellini. Bellini painted on another wall **a vat of red wine with some Bacchanals around it**, and Satyrs, musicians, and other men and women, all drunk with wine [. . .] which work was in truth executed and colored with great diligence, insomuch that it is one of the most beautiful pictures that Gian Bellini ever painted, although in the manner of the draperies there is a certain sharpness after the German manner (. . .) because he imitated **a picture by the Fleming Albrecht**

> **Dürer**, which had been brought in those days to Venice and placed in **the Church of S. Bartolommeo**, a rare work and full of most beautiful figures painted in oils.

In particular, the phrase *stories of Æneas, Mars, and Venus* is difficult to interpret. Based solely on the de Vere text, it is unclear whether these names refer to three separate existing paintings, a single lost multi-scene work, or commissions by Dosso that Vasari later conflates with works by Titian and Bellini. The steps outlined above do not resolve this issue without external confirmation, as the reference cannot be clearly identified from the context alone.

To resolve this, we consulted the Wagenbach critical edition of the biography (Vasari 2005). The footnotes helped us identify that the description refers to a full cycle of mythological canvases commissioned from multiple artists, several of whom are identifiable in modern reference works. Cross-referencing the editorial apparatus with current Wikidata entries yielded various annotation decisions, of which we want to exemplify the following:

- **Generic Mention:** *stories of Æneas, Mars, and Venus* – The collective phrase functions as a program description rather than a reference to any single recoverable artwork. Because it denotes a set of commissions without individuating them, it is annotated as a *Generic Mention* with no entity identifier assigned.
- **Implicit Artwork Mention:** *Aeneas* [in the Elysian Fields] – Dosso Dossi's *Aeneas in the Elysian Fields*, one of the surviving canvases of the *Camerino* program, is recoverable from the narrative context and identifiable in Wikidata. The Wagenbach apparatus confirms this. The mention is annotated as *Implicit Artwork Mention* and linked to the corresponding Wikidata entity.

This example shows our commitment to creating a gold-standard GT. When the de Vere translation does not provide sufficient context to clarify meaning and the referent can likely be found in established scholarly sources, we use the Wagenbach edition and relevant art-historical catalog entries as secondary evidence. This approach matches the broader goals of this dissertation, which treats annotation decisions as interpretive acts. The GT records a cumulation of these choices (see Section 2.4).

Ground truth post-annotation report. The annotation process described earlier produced a GT corpus with 2,422 mention instances, distributed across 232 paragraphs from 16 artist biographies. Table 7.11 shows the breakdown for each biography, including the number of paragraphs, total mentions, unique entities, Wikidata-linked entities, and OOKB entities. The corpus was exported from INCEpTION as UIMA CAS JSON, then converted to the JSON format used in our pipeline (see Table 7.9). Mentions of the same artwork across different paragraphs

are assigned the same entity identifier, so coreferent mentions are grouped by identity rather than by how close they appear in the text. During quality control, we identified and corrected headless coreferential chains by merging them with their correct antecedents. In the final corpus, every coreferent mention has a resolvable `refers_to` pointer to an antecedent within the same biography. This correction reduced the corpus from an earlier count of 2,439 mention instances to the 2,422 reported here. The reduction matters for downstream reading of the populated KG, since the deployed graph (Section 7.5, Table 8.4) was populated *before* this correction and therefore still carries the older 2,439-mention GT layer. The figures in this chapter are the corrected, authoritative ones (2,422 mentions, 790 implicit / 220 explicit, 848 entities) and are the version used for all recognition and linking evaluation in Chapter 9. The human-annotated layer of the populated graph in Chapter 8 (2,439 mentions, 811 / 221, 852 GT artwork entities) instead reflects this pre-correction GT, which the canonical build — *population Run 2*, on which every KG metric is computed (Section 8.3) — still carries over unchanged from the earlier Run 1, and which a future re-deployment will reconcile to the corrected corpus. The mismatch is confined to the human-annotated GT layer and leaves the ObliquER recognition layer untouched.

By analyzing these annotations, we found that explicit artwork mentions are relatively few because Vasari often uses anaphoric expressions or descriptive phrases once artworks have been introduced, and even aside from re-appearing artworks, implicit mentions dominate the GT corpus. Across the 2,422 surface references annotated in the corpus, explicit artwork mentions account for only 220 instances (9.1%), while implicit mentions contribute 790 (32.6%), coreferent mentions 1,080 (44.6%), and generic mentions 332 (13.7%) (Table 7.11). This quantifies the corpus condition that RQ2 presupposes, with the set of implicit mentions correspondingly – indeed, inherently – large. The 1,010 explicit and implicit mentions (21.8% explicit, 78.2% implicit) form the evaluation set for ObliquER. Figure 7.27 shows this breakdown for each biography, arranged by Vasari volume.

After resolving the data, the 956 entity clusters from individual biographies are reduced to 848 unique artwork entities at the corpus level. This reduction happens for two reasons. Entity IDs are specific to each biography, with each biography numbering its mentions starting from `0001`. When populating the KG with GT triples, we aim to avoid duplicates across the entire dataset. Instead, we use a resolved knowledge-base identifier from either the Wikidata QID or the Viewsari internal KB. When an artwork or architectural site appears in more than one biography, the shared identifier combines what would otherwise be counted as separate entities. For example, the Sistine Chapel is mentioned in the biographies of Botticelli, Ghirlandaio, Pollaiuolo, and Rosselli. Similarly, Santa Maria del Fiore appears in five biographies (Brunelleschi, Cimabue, Ghirlandaio, Giotto, Verrocchio). Of the 848 resolved entities, 310 are linked to Wikidata identifiers, and 538 are OOKB entities without a stable external identifier. The resulting split, about 63.4% OOKB and 36.6% Wikidata, highlights the limited coverage of Wikidata for minor commissions, lost works,

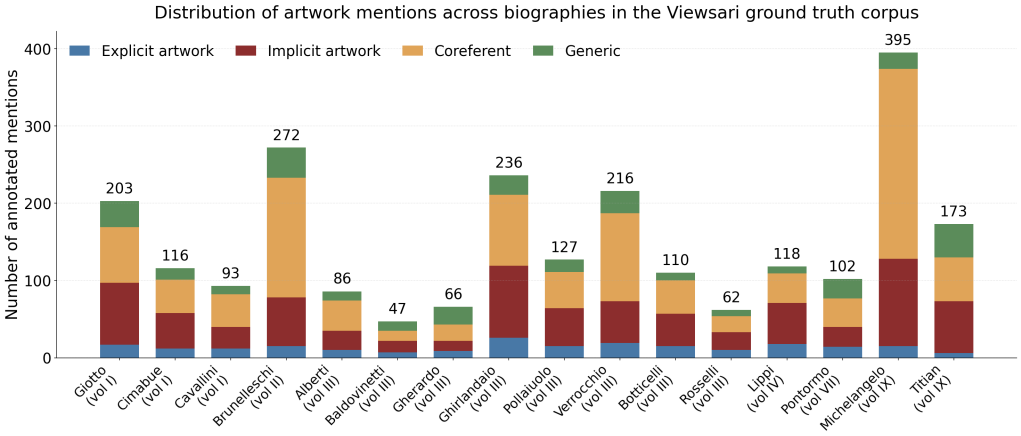


Figure 7.27: Distribution of annotated artwork mentions across the 16 biographies of the Viewsari GT corpus, ordered chronologically by Vasari volume (in parentheses below each label). Each bar is stacked by mention type – explicit artwork, implicit artwork, coreferent, and generic – and annotated with the per-biography total. The corpus-wide skew toward coreferent and implicit mentions is visible across biographies of all periods, with Michelangelo's biography ($N = 395$) and Brunelleschi's ($N = 272$) standing out both in absolute counts and in the relative weight of coreferent chains characteristic of Vasari's longer narratives.

and works with uncertain attribution. This imbalance is not a flaw in the annotation method but reflects the realities of the field. It also sets the context for our hypothesis that, if a large proportion of the artworks Vasari discusses are not in structured KBs, any extraction process limited to linkable entities will underestimate both recall and the interpretive scope. ObliquER's ability to create records for artworks not in KBs by assigning internal identifiers and keeping textual descriptions when Wikidata links are missing is therefore treated as a core requirement, not just a backup.

This GT is the main evaluation resource for the experiments described in Chapter 9. It is the first annotated dataset for recognizing implicit artwork entities in Renaissance biographical prose. The dataset was built following the annotation guidelines, sampling procedure, and Wagenbach-based disambiguation strategy explained earlier. Both the dataset and the annotation guidelines are available in the project repository.[23]

Development split. Few-shot prompts can lead to inflated evaluation results if any example paragraph appears in both the prompt and the test set. To prevent this, we added a script that creates a separate development partition from the GT and saves it. This same file is then referenced in the development field of every Hydra dataset configuration (see Section 7.4.8.1), so both inference and

[23] `https://github.com/ISE-FIZKarlsruhe/ObliquER/tree/main/data/viewsari`.

Table 7.11: Per-biography breakdown of the Viewsari GT corpus by mention type. *Bios* stands for the artist biography; *Paras* for the number of paragraphs per biography. *Exp.*, *Imp.*, *Coref.*, and *Gen.* count, respectively, explicit artwork mentions, implicit artwork mentions, coreferent mentions, and generic mentions. *Mentions* is the row sum across all four types. *Entities* counts unique artwork `entity_id` clusters per biography; *LINKED* and *OOKB* report the knowledge-base coverage split between Wikidata-linked and OOKB entities. † The *Entities*, *LINKED*, and *OOKB* totals are raw per-biography counts taken before corpus-level deduplication (the 956 clusters reduce to 848 unique entities). Per biography, *LINKED* + *OOKB* can be smaller than *Entities*, because some clusters are left unresolved at annotation time, carrying neither a Wikidata link nor an OOKB flag; the 18 such clusters across the corpus account for the difference ($956 = 396 + 542 + 18$) and are assigned a resolution during deduplication.

Bios	**Paras**	**Exp.**	**Imp.**	**Coref.**	**Gen.**	**Mentions**	**Entities**	**LINKED**	**OOKB**
Giotto	15	17	80	72	34	203	90	39	51
Cimabue	14	12	46	43	15	116	58	22	35
Cavallini	7	12	28	42	11	93	40	13	26
Brunelleschi	16	15	63	155	39	272	77	28	47
Alberti	11	10	25	39	12	86	35	15	20
Baldovinetti	8	7	15	13	12	47	22	6	16
Gherardo	5	9	13	21	23	66	21	8	12
Ghirlandaio	19	26	93	92	25	236	117	51	60
Pollaiuolo	14	15	49	47	16	127	64	12	52
Verrocchio	17	19	54	114	29	216	71	27	41
Botticelli	14	15	42	43	10	110	55	39	14
Rosselli	4	10	23	21	8	62	33	16	17
Lippi	9	18	53	38	9	118	70	20	50
Pontormo	15	14	26	37	25	102	40	16	23
Buonarroti	49	15	113	246	21	395	90	41	49
Titian	15	6	67	57	43	173	73	43	29
Total	**232**	**220**	**790**	**1,080**	**332**	**2,422**	**956**†	**396**†	**542**†

evaluation are limited to the test partition. The development partition includes every paragraph used as a demonstration in any prompt variant, allowing all variants to be evaluated on a single shared test partition without recalculating exclusions for each variant (see Section 7.4.8.4).

The prompt structure used in ObliquER derives systematically from the ontology requirements engineering presented in Section 7.3.1 and Section 7.3.2. The role descriptions in the system prompt are grounded in the user personas, developed during requirements analysis, and the task formulations operationalize the CQs that drive the Viewsari ontology (Chapter A). This moves the requirements to inference-time prompt artifacts, in contrast to the largely undocumented prompt construction reported in much of the LLM-based extraction literature (Dagdelen et al. 2024, Zhang et al. 2025e). The role description, in particular, is transcribed directly from a synthesis

of the personas to preserve the requirements-engineering grounding through to the prompt level. In Chapter 9, we investigate prompt sensitivity through ablation studies (Section 9.4.6), varying components of the pipeline architecture at the prompt level. However, we note that exhaustive testing of prompt variants is combinatorially infeasible and is not standard practice in the field; recent work on prompt sensitivity (Mizrahi et al. 2024, Sclar et al. 2024) establishes ablation along principled dimensions as the methodological standard for characterizing the influence of prompt formulation on model output. The ablation results accordingly report the effects of each such variation to characterize prompt sensitivity rather than provide an exhaustive list.

7.4.8 Entity recognition with ObliquER – spotting the unnamed

The ObliquER pipeline, illustrated in Figure 7.28, consists of a dynamic-chunked recognition stage that types in-text spans as explicit, implicit, or coreferent, and a candidate-generation-and-ranking linking stage that maps the resulting mentions to Wikidata entities or assigns them an OOKB label. The present section addresses the first stage; Section 7.4.9 turns to the second. We begin with the prompt design rationale for recognition, then describe the modular template infrastructure shared across both stages, and finally introduce the three recognition prompt variants evaluated in Section 9.4.

7.4.8.1 Architecture of the recognition pipeline

Within ObliquER, entity recognition is formalized as a generative extraction task and operationalized using LLMs. The pipeline processes sequences of paragraphs, specifically prior, current, and subsequent segments, from biographical texts, and produces structured JSON records for each detected mention, conforming to the GT annotation schema introduced in Section 7.4.6.2.

The architecture of the recognition pipeline is composed of the following principal components:

Corpus preparation. Prior to prompt assembly, the raw volume files are preprocessed to segment Vasari's text at the paragraph level; the output is a single file per volume with the different paragraphs.

Prompt assembly. For each input, we assemble prompts based on the prompt parts elaborated in Table 7.12, including the role, entity definitions, task instructions, paragraph chunks, etc.

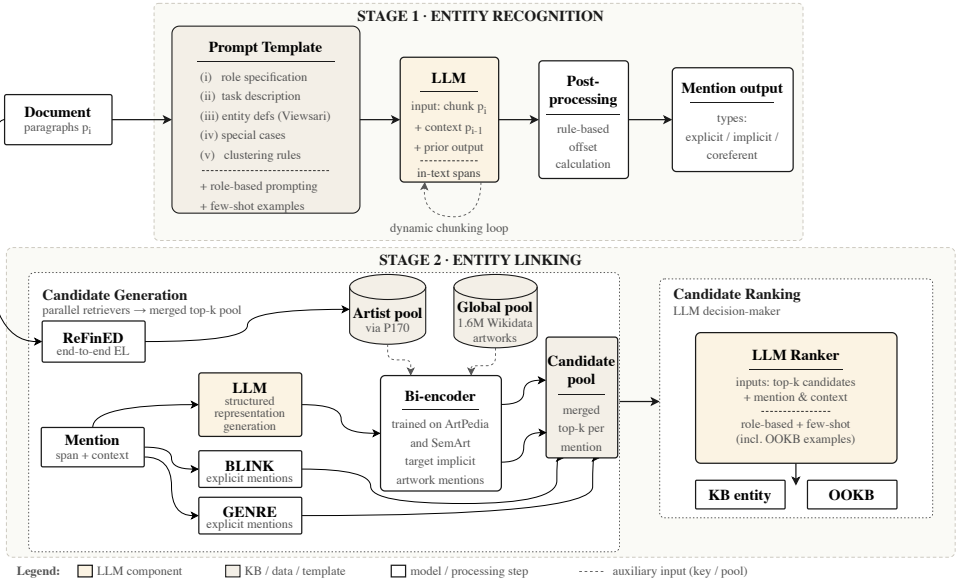


Figure 7.28: The ObliquER pipeline. Stage 1 performs dynamic-chunked mention extraction with an LLM, typing spans as explicit, implicit, or coreferent. Stage 2 merges candidates from ReFinED, BLINK, GENRE, and an ArtPedia/SemArt-trained bi-encoder, then an LLM ranker selects a KB entity or assigns OOKB.

Dynamic chunking. The biographical text is segmented into overlapping paragraph sequences, a strategy that preserves contextual information while respecting the constraints imposed by the model's context window. This approach provides the coreference context necessary to resolve short-range anaphoric expressions (e.g., *it*, *the work*) while avoiding model instability associated with very long, unstructured context windows (Sheng et al. 2025, Hosseini et al. 2025). Prompts are processed in document order, and the raw string response from each inference call is assigned directly to the prior-output slot when rendering the subsequent prompt. By injecting the prior output, the process retains all information generated by the model that may be relevant for downstream coreference resolution.

A consequence of this chunking strategy, limited to only the previous paragraph and its output, is that coreferent mentions whose antecedent lies more than one paragraph away cannot be resolved within the recognition stage alone. In the future, we aim to resolve long-range chains during entity cluster unification.

Large language model inference. For each paragraph, the assembled prompt is submitted to the LLM via an endpoint (see Section 7.4 for details on models and versions). Within a biography, inference is strictly sequential. For each paragraph, the slot reserved for the prior

output is populated with the raw model response to the immediately preceding paragraph (or, for the first paragraph of a biography, with a fixed placeholder string), while the role specification is supplied separately as the system message. Each model response is persisted alongside its source prompt. To accelerate large extraction campaigns, multiple biographies are processed concurrently while paragraphs within a single biography remain sequential, preserving the prior-output dependency that enables cross-paragraph coreference resolution. The model returns the mention records produced for the paragraph.

Post-processing. The raw output from the model is subsequently subjected to post-processing, which normalizes mention types, verifies the presence of surface forms, and computes character-level offsets necessary for integration into the Viewsari KG. We post-process the output to harmonize type labels that were mutated by the LLM (e.g., producing `IAM` instead of `implicit artwork mention`). Moreover, we unify entity clusters by matching linked entities across paragraph boundaries. A notable limitation is that this is not possible for OOKB-detected entities. Therefore, duplicate entries of OOKB entities are possible.

Inference orchestration with Hydra. The full inference loop is parameterized through a Hydra [24] configuration tree that composes a *model* configuration – which fixes the inference target to `openai-gpt-oss-120b` – with a *dataset* configuration. Each prompt variant (few-shot baseline, CoT, ontology-guided, and the ArtPedia transfer setting) has its own such configuration. A flag activates the sliding window described in Section 7.4.8.1; an optional development-split specification restricts inference and evaluation to a held-out development partition of the GT, which is the mechanism we use to keep all few-shot example paragraphs strictly disjoint from the evaluation set described in Chapter 9.

7.4.8.2 Recognition prompt architecture

Both stages of the ObliquER pipeline (recognition, then linking; see Section 7.4.9) are based on structured prompts assembled from modular components during inference. The prompt design is calibrated against the prompting-mode taxonomy introduced in Section 5.1.3: the modular content–structure separation described below is a deliberate response to the observation that prompt-engineering decisions function as provenance objects in their own right (Section 5.1.3.3),

[24] Hydra framework for complex applications: `https://github.com/facebookresearch/hydra`.

and the three production variants are positioned across the zero-shot, few-shot, and reasoning-scaffold axes surveyed in that section. This section describes the prompt design rationale, the template infrastructure, and the three prompt variants evaluated.

Table 7.12 summarizes which content layers are activated across the recognition prompt variants.

Table 7.12: Content layers activated across the ObliquER recognition prompt variants (Few-shot, Ontology-guided abbreviated as Onto., CoT) and the four ablation conditions A1–A4. ✓ = layer present; – = layer absent. A1 collapses recognition, linking, and coreference into a single prompt; A2 strips the domain-specific content layers down to a generic NER role and task; A3 keeps the ontology-typed entity definition of the ontology-guided variant but drops the KG modeling-example block and supplies no demonstrations; A4 varies only the demonstration count over the ontology-guided template. The empirical comparison of the seven configurations is reported in Section 9.4.6.

Content layer	Few-shot	Onto.	CoT	A1	A2	A3	A4
Expert role description	✓	✓	✓	✓	–	✓	✓
Entity definition / taxonomy	✓	✓	✓	✓	–	✓	✓
Task instructions	✓	✓	✓	✓	✓	✓	✓
Special-case rules	✓	✓	✓	✓	–	✓	✓
Ontology alignment	–	✓	–	–	–	✓	✓
KG modeling example	–	✓	–	–	–	–	✓
Reasoning scaffold	–	–	✓	–	–	–	–
Linking subtask in same pass	–	–	–	✓	–	–	–
Few-shot examples (count)	1	2	1	1	0	0	{1,3,5}

Modular template design. All prompts are constructed from two complementary artifacts:

Static content files contain the domain-specific textual building blocks: role descriptions, entity definitions, task instructions, special-case rules, clustering rules, output format specifications, and few-shot examples. Each prompt variant maintains its own static content file, so that changes to one variant do not propagate to others.

Templates define the structural layout of the final prompt. They reference the static building blocks by name and inject dynamic, per-paragraph context at render time. We use this to separate structure from content.

We enforce this separation to add flexibility to the approach where no domain-specific content is hard-coded in the components; a strict content–structure boundary allows modifying which task the model should perform by editing only static content.

Dynamic variables. At inference time, the renderer injects two pieces of paragraph-level information into each recognition prompt, namely the identifier and full text of the current paragraph, and the identifier, text, and raw model response of the immediately preceding paragraph for additional context. The target entity class for recognition (*Artwork*) is fixed in the static content rather than supplied per paragraph. Recognition prompts are pre-rendered into one prompt instance per paragraph per volume before an inference call.

7.4.8.3 Further details on prompts for entity recognition

For ObliquER, we crafted prompt templates composed of building blocks to dynamically address specific aspects relevant to entity recognition, the so-called 'instructions'. The approach was mainly influenced by our annotation process for the Viewsari GT, as discussed in Section 7.4.4. The building blocks for the prompts are, in order of appearance: (i) a role specification, followed by (ii) a task description, (iii) an entity definition and its ontology grounding, (iv) a KG modeling example (ontology-guided variant and its A4 descendants only), (v) special cases, and, in the CoT variant, (vi) a standalone clustering-rules block (folded into the special-cases block in the ontology-guided variant and absent from the few-shot baseline). Table 7.12 summarizes which of these blocks each variant uses; the remainder of this subsection describes each block in turn.

(i) Role specification. The first part of each prompt is the role specification, which defines the role as that of a domain expert, inspired by the personas we developed in Section 7.3.1 and Section 7.3.2. Role-based prompting (Chen et al. 2025a) directs the model to apply a domain-specific lens to the input text, improving sensitivity to art-historical referential conventions.

> You are an expert in historical text analysis with a strong background in interpreting biographical texts from the Renaissance period, especially Giorgio Vasari's *Lives of the Most Eminent Painters, Sculptors, and Architects* (1555/1568). You are familiar with the work's style, language, and context.

This static role description already conveys several contextual signals, including the disciplinary framing of historical text analysis, the temporal-geographic register of the Renaissance, and the specific mention of Vasari's *The Lives*. The subsequent prompt blocks – entity definitions and the mention taxonomy (iii), special-case rules (iv), clustering rules (v), and, in the ontology-guided variant specifically, an ontology-typed entity definition paired with a small KG modeling example (Section 7.4.8.4) – extend this signal by constraining what counts as an artwork and how the recognition decisions are to be structured.

In evaluating prompt templates, we hypothesized that specifying a domain-expert role would be insufficient to reliably recognize artwork entities. To address this limitation, we introduced an input-dependent sentence that explicitly identifies the artist under analysis. This intervention is grounded in empirical evidence that artist-constrained candidate filtering improves EL performance (see Section 7.4.9.4). If knowledge of the artist improves disambiguation during linking, then incorporating this information at the recognition stage should similarly enhance accuracy. Accordingly, we parameterized the CoT recognition variant to include the artist as contextual input.

The underlying mechanism, though, mimics the principle of topical coherence, which models how human annotators might disambiguate a mention referentially through the creator's identity when the model is supplied with additional context, such as the paragraph under analysis being part of a specific artist's biography.

When we take the example from Section 7.4.7 and the mention of *those ornaments worn on the head by the girls of Florence*, an anchor to Ghirlandaio and the context of his family name's origin based on these garlands, the model can resolve the mention not from the paragraph text alone, but with additional contextual information. Knowing that we are reading the biography of a specific artist, possibly one that we know, shapes our interpretation of *those ornaments worn on the head by the girls of Florence* differently. The observation that mentions within a document are far more likely to refer to entities related to the document's central topic than to unrelated entities with the same surface form (Milne and Witten 2008, Globerson et al. 2016) lays the foundation for our approach; related literature documents that supplementary metadata as part of the input context consistently improves recognition performance for low-frequency and ambiguous entity types (Zhou et al. 2024c, Chen et al. 2025a).

Yet, informal comparisons during prompt development did not reveal a measurable improvement (F1 accuracy did not reflect a benefit attributable to involving an artist's name); a controlled evaluation of this block is left to future work (Section 12.3). Correspondingly, we did not incorporate this block as a static component of the prompts for the few-shot and ontology-guided variants, both retaining the biography-agnostic role description quoted above. The published architectural ablation (Section 9.4.6, A1–A4) does not examine this information on the artist's biography, since A1–A4 vary across other dimensions (tasks, domain content, ontology layer, example count). The disambiguation argument that motivated the experiment is nonetheless instructive and remains the basis for the artist-constrained retrieval that pays off at the linking stage (cf. Section 7.4.9).

(ii) Task description. We designed this task block to instruct the model to extract all mentions of artwork entities from the input. The subsequent block then provides explicit definitions and examples for each mention type.

The full task instruction reads as follows:

> Your task is to extract all `ARTWORK` entities mentioned in the input text, including any creative product described or implied as an artwork, even if implicit. This includes longer descriptive passages or scenes from artworks that might be referenced by their thematic content rather than a formal title.
>
> Include the following textual references to artworks:
>
> - Explicitly named artworks (proper names, established titles, or traditional identifiers such as "La Gioconda"). Select the full title span naming the artwork.
> - Implicitly referred artworks (descriptive phrases identifying a specific, recoverable work). Select the full descriptive phrase.
> - Coreferent mentions (subsequent references to an artwork already introduced – pronouns, demonstratives, definite noun phrases, or descriptive repetition).
>
> You are also given the previous paragraph's output. Artworks in the current paragraph may refer to works introduced in the previous paragraph; link them using `refers_to`.
>
> In the output, give the `mention_id` of the antecedent explicit or implicit mention under `refers_to` for all coreferent mentions. Coreference chains can span across paragraphs.

The multi-mention coreference rule embedded in the task block encodes our design decision for cross-paragraph coreference behavior. In ObliquER, the model is instructed to adhere to a simulated sliding-window architecture, where, within a self-contained paragraph with no prior context, the first mention of an artwork is always annotated as *Explicit* or *Implicit* and never as *Coreferent* because there is no antecedent to point to.

All subsequent mentions of the same artwork within a paragraph must reference the same antecedent identifier, regardless of their proximity to the initial mention. Generic head nouns, including *work* or *it*, are annotated as coreferent only when they refer to a recoverable antecedent; otherwise, they are excluded from annotation. This protocol is designed to prevent the model from inferring coreference when only implicit artwork mentions are present. Coreferent mentions

are excluded from evaluation metrics and serve solely to keep these references traceable in the source text.

The sliding-window architecture imposes a strict limit on the length of coreference chains that can be resolved during recognition. When a coreferent mention's antecedent spans multiple paragraphs, it may fall outside the model's context window and cannot be linked at recognition time. This constraint arises from the limited context windows of open-source LLMs relative to the length of full biographies. We address this limitation in the EL phase, where mentions are reconciled using knowledge-base identifiers. Clusters produced independently during recognition are assigned the same identifier via linking, enabling post-processing to merge them into a single canonical entity.

For entities not present in the KB, long-range coreference chains remain unresolved unless the entity labels and contextual descriptions generated during recognition are sufficiently similar to trigger heuristic cluster merging in post-processing. This limitation is inherent to the current architecture and is examined in detail in Section 7.4.9 and Chapter 9.

(iii) Entity definitions and their Viewsari ontology grounding. The entity definition block grounds the extraction task in the Viewsari ontology (see Section 7.3), but the extent of this grounding depends on the prompt variant. The few-shot baseline (and the monolith ablation A1) embeds a domain-flavored description of what counts as an artwork in Vasari's text – enumerating artwork forms by their everyday names (panel paintings, altarpieces, frescoes, statues, churches, drawings, tapestries, and so on) but without naming Viewsari class descriptions. The ontology-guided variant (and its ablation descendants A3 and A4) replaces this with a fully ontology-typed definition. The CoT variant uses a compact summary form:

> An Artwork is a creative or artistic product. This includes paintings, sculptures, architectural constructions, altarpieces, or decorative artworks. It needs to refer to a specific artwork, not to a category or a process that describes a person's interest or talent.

Each of the three forms employs a positive specification of artwork mentions, paired with a negative clause that excludes process descriptions and generic mentions. Positive descriptions of concepts such as *paintings, sculptures, architectural constructions, altarpieces, or decorative artworks* enumerate the ontological subclasses of `viewsari:artwork`, whether or not the variant names that class explicitly, so that the conceptual grounding is preserved across all three production prompts. The negative clause excludes mentions that refer to a category or to a process describing a person's interest or talent, which constitute the most frequent class of false positives observed during prompt development. Generic phrases such as *his skill in fresco*, *a talent for sculpture*,

or *his manner of painting*, which superficially resemble artwork descriptions but in fact refer to an artist's practice rather than to a specific object, were cases we noticed during the annotation process. The boundary between generic and implicit is not always apparent, so the negative clause intends to make this distinction explicit to the model. The contrast between this implicit and explicit form of ontology grounding is one of the dimensions tested by the architectural ablation of Section 9.4.6.

(iv) Knowledge-graph modeling example. The ontology-guided variant (and its ablation variants) also includes a Viewsari KG modeling example. As such, a dedicated block provides a short demonstration showing how an artwork mention extracted by the model is materialized as a set of RDF triples in the populated Viewsari KG. The purpose is to expose the model to the precise schema that the output will fall into downstream, which aims to reduce type-naming drift across runs (a controlled measurement of this effect is left to future work, Section 12.3), and to merge, in a single artifact, recognition as not merely classification but as the first step of a KG population pipeline.

(v) Special cases. We added a section focused on recurring error patterns and boundary cases that we identified while developing this prompt structure and manually annotating the GT in Section 7.4.4. One of the main issues is distinguishing between implicit artwork mentions and general artistic references, as explained in Section 7.4.6.4. Another common mistake is the way iconographic subjects are annotated with implicit mentions. For example, annotating *the Last Supper* as an *Explicit* mention instead of treating *a panel for the friars of Santa Croce representing the Last Supper* as the *Implicit* mention span. We also found that coreferences across paragraph boundaries pose a major challenge (see Section 7.4.8.1), which is the main structural limitation of the sliding-window architecture. The special-cases section clarifies that the first mentions in a paragraph are not always new entities, so it is important to check previous output before assigning a type.

Finally, we want to highlight that architectural artworks, especially buildings, are challenging to annotate. The text often alternates between describing a building as a creative achievement, thereby making it an *Artwork* entity, and referring to it as a location or institution, often in relation to other artworks. In these cases, the building is not considered the artwork mentioned.

(vi) Clustering rules. All three recognition variants expect the model to maintain consistent entity clusters across mentions, meaning that every mention of the same artwork within the current paragraph shares a single `entity_id`, assigned at the first mention and reused for all subsequent coreferent mentions (resolved then in the postprocessing described in Section 7.4.8.1). This is

consistent with the Viewsari ontology and the KG population, in which there is exactly one node per artwork, and additional related nodes correspond to textual references to that artwork.

The variants differ only in where this is encoded in the prompt. The CoT variant has a dedicated clustering-rules block that follows the special-cases block in the rendered prompt. The ontology-guided variant folds this into the special-cases block. The few-shot baseline implicitly propagates this rule through the worked example embedded in the template. In all cases, the resulting output schema is identical.

7.4.8.4 Recognition prompt variants

The preceding subsections established an abstract framework for prompt building blocks. As previously noted, the manifestation of these elements is contingent on the selected prompt variant. To clarify this dependency and prepare for the subsequent evaluation, we introduce the three recognition-prompt variants that were implemented. Each variant employs the same modular infrastructure but diverges in its inference strategy and alignment with the underlying ontology.

Few-shot (standard) prompts. The baseline variant presents the task instructions, a single extended example from the Ghirlandaio biography, the previous paragraph context, and the current paragraph.

This is our first prompt for the few-shot strategy (Brown et al. 2020, Phoenix and Taylor 2024, Ji and Kong 2024), surveyed in Section 5.1.3.2, which is adopted to expose the model to a diverse set of mention types without requiring fine-tuning. We hand-pick a few-shot example paragraph from the GT, which is later excluded from the evaluation that includes a variety of artwork and mention types, and give the model the entire example input passage and the corresponding annotations.

The few-shot template includes the annotated output exactly as it should appear. Each record indicates to the model which fields to produce for a mention, including a paragraph-local identifier, mention type, surface form, paragraph-local entity-cluster identifier, antecedent pointer for coreferent mentions, and an in-text span annotation that marks the surface form in its context with double-angle-bracket markers. The first three records in the demonstration are an implicit artwork mention, a second implicit mention nested inside it, and a coreferent resumption. There are twelve mention records in total, covering all three main mention types and demonstrating the cross-mention coreference rule.

```
{
 "paragraph_id": "251",
```

```
 "mention_id": "m_0035",
 "type": "implicit artwork mention",
 "surface_form": "a panel for their high-altar",
 "entity_id": "entity_0023",
 "refers_to": null,
 "in-text span annotation": "the Frati Ingesuati he painted
   <<a panel for their high-altar>>, with certain Saints kneeling"
},
{
 "paragraph_id": "251",
 "mention_id": "m_0036",
 "type": "implicit artwork mention",
 "surface_form": "their high-altar",
 "entity_id": "entity_0024",
 "refers_to": null,
 "in-text span annotation": "he painted a panel for
   <<their high-altar>>, with certain Saints kneeling"
},
{
 "paragraph_id": "251",
 "mention_id": "m_0037",
 "type": "coreferent",
 "surface_form": "this work",
 "entity_id": "entity_0023",
 "refers_to": "m_0035",
 "in-text span annotation": "and other saints. For
   <<this work>> Domenico truly deserves praise"
}, ...
```

We choose examples that contain all three mention types (*Explicit*, *Implicit*, *Coreferent*). Simultaneously, we ensure that at least one instance of each recurring error pattern identified during prompt development is included in the few-shot examples.

A focal point is the application of dynamic chunking, which prompts LLMs to process paragraph sequences by providing the preceding paragraph's output along with the current paragraph's text as additional context (Sheng et al. 2025).

Ontology-guided prompts. The ontology-guided variant extends the standard few-shot prompt with an additional modeling-example block that shows how an extracted mention is represented in the Viewsari KG. This block presents a complete downstream RDF serialization for a mention drawn from a *The Lives* paragraph, illustrating the mapping from extraction output fields to ontology classes and properties.

The ontology-guided entity definition also explicitly uses the Viewsari ontology class names (e.g., *explicit artwork mention*, *implicit artwork mention*, *coreferent*, *generic mention*), thereby aligning the model's output vocabulary with the downstream schema. This variant includes two extended few-shot demonstrations which jointly cover all four mention types (including *generic mention*), Wikidata-linked architectural works, and more complex coreference chains. Conceptually, this variant operationalizes the framing of generative IE as annotation production discussed in Section 5.1.3.3: by supplying the model with the ontology classes that the downstream graph will use, the prompt elicits outputs that can be ingested as annotations without an intermediate schema-mapping step.

Ablation A1: monolith prompt. The first ablation tests whether decomposing the pipeline's different extraction parts – recognition and a separate EL pass over the candidate-generation stage – yields measurable gains over a single end-to-end prompt. In the monolith variant, the three pipeline stages of recognition, EL, and coreference resolution are collapsed into one prompt. Consequently, the model infers for each artwork mention one of the four mention types and the corresponding linked cluster, directly mapping them to a Wikidata identifier derived from its own parametric knowledge of the world. Because the linking step in this condition does not consult any other information described in Section 7.4.9.4, the variant also functions as a lower-bound estimate of how much the retrieval stage contributes when added on top.

Ablation A2: no domain-specific content. The second ablation tests whether the domain-specific content layer – expert role framing, the artwork-mention taxonomy, the few-shot example(s) – contributes measurably to performance. The structural template, the output schema, and the four-way mention vocabulary are kept identical to the few-shot baseline; only the substantive content is stripped. Specifically, the expert role description is replaced with a generic assistant role, the detailed entity definition is collapsed into the bare instruction *"identify artwork mentions in the input text"*, the special-cases block is removed in full, and the few-shot example block is rendered empty so the model receives only the task description and the current and previous paragraphs. This condition isolates the contribution of the content layer relative to the structural template and is the variant that most directly answers whether the modular block decomposition described in Section 7.4.8.3 justifies its complexity.

Ablation A3: zero-shot ontology prompt. The third ablation is a zero-shot approach, positioned against the zero-shot extraction literature surveyed in Section 5.1.3.1. It preserves the ontology as the structural backbone, making it a variant of ontology-guided prompt templates. In contrast to the native prompt design of these in Section 7.4.8.4, the zero-shot variant supplies no annotated demonstrations and no KG modeling example, isolating the contribution of demonstrations from that of ontology grounding.

Ablation A4: ontology with variable example counts. The fourth ablation, as in A3, uses the ontology-guided template and only varies the number of examples in the demonstration list for the example $k \in \{1, 3, 5\}$. The example sets are statically embedded and drawn verbatim from the GT corpus, then excluded from the evaluation. Because the production ontology-guided variant uses $k = 2$, in our evaluation, Table 9.18 reports $k \in \{0, 1, 2, 3, 5\}$ where the $k = 0$ row shows A3 and the $k = 2$ row shows the results of the native ontology-guided prompt template.

The empirical comparison of few-shot, CoT, and ontology-guided variants is reported in Section 9.4.1. The best performing variant, in this case, ontology-guided at $k = 2$, is additionally re-run under a different inference model in the cross-model evaluation of Section 9.4.5, where the ontology-guided template is held fixed and *both* the inference decoder (`openai/gpt-oss-120b` vs. `google/gemma-4-31b-it` (Gemma Team, Google 2026)) and the demonstration count ($k \in \{0, 1, 2, 3, 5\}$) are varied. A1–A4 and the cross-model evaluation are complementary in that the former holds the decoder fixed and varies the prompt structure, while the latter holds the prompt structure fixed and varies both the decoder and the demonstration budget along a single axis.

7.4.8.5 Chain-of-thought recognition prompting

In addition to the standard few-shot prompt and the ontology-guided variant, ObliquER implements a CoT recognition prompt that adds a structured reasoning step between input and output. This variant augments the standard prompt with a structured reasoning scaffold. The models employed here perform latent reasoning regardless of the prompt variant; what distinguishes the CoT variant is therefore not the presence of reasoning but its elicitation through an explicit, instruction-specified scaffold whose intermediate steps are rendered into the structured output. The model is instructed to reason through diagnostic questions for each candidate mentioned:

1. Is this phrase referring to a specific, recoverable object – or to an artistic practice, talent, or generic category?

2. Is this the first introduction of this artwork in the current paragraph, or does it resume an entity already introduced?
3. What contextual signals support identification (creator, location, commissioner, iconographic subject, medium, date)?
4. Is the referent recoverable from context, or genuinely ambiguous?

The appended few-shot demonstration is drawn from the same paragraph as the standard few-shot baseline and is supplemented with a reasoning trace.

This variant additionally extends, as discussed in Section 5.1.3.3, the focus to self-consistency checks by the model via intermediate reasoning steps before producing a final answer (Wei et al. 2023). We deem this particularly relevant to both implicit and explicit artwork mentions, as identifying such mentions is inherently inferential.

As established in Section 5.1.3.3, CoT prompting does not consistently improve recognition performance. The empirical performance of our CoT variant relative to the standard few-shot prompt is reported in Chapter 9. For the Viewsari KG, the CoT outputs serve as provenance objects, providing additional context for the provenance statements modeled using PROV-O and the Web Annotation Ontology. Although every variant relies on a model that reasons internally, only the CoT variant elicits that reasoning through instructions and renders it into a recoverable trace; the models' intrinsic reasoning remains unobservable and cannot serve as a provenance object. We store the results in the repository and, in future extensions, plan to integrate these trails into the Viewsari KG. As argued in Section 5.1.3, when reasoning is preserved alongside the extracted mention, the KG can be populated not only with *what* was extracted and from where by which model, but *why* the model concluded that a given phrase constitutes an artwork reference at this point in time.

7.4.9 Entity linking with ObliquER – a sea of unknowns?

EL, the second stage of the ObliquER pipeline (see Figure 7.28), follows the standard two-step retrieve-then-rank paradigm (Logeswaran et al. 2019, Wu et al. 2020), and is positioned within the generative EL literature surveyed in Section 5.1.7: rather than treating linking as a closed-vocabulary classification problem, the ranker is an LLM that operates over a retrieved candidate pool and is permitted to emit `NIL` when the correct referent is absent from the KB, which makes the long-tail and OOKB regime of Section 5.1.8 a first-class concern of the design rather than a downstream failure mode. Exhaustively scoring every mention against the full Wikidata artwork index would be infeasible at inference time, so a candidate-generation stage first narrows the search

space to a small, recall-oriented pool that the ranker then scores precisely against the mention and its surrounding context. Four retrievers run in parallel during candidate generation, chosen so that their union covers both explicitly named and implicitly described artworks: ReFinED (Ayoola et al. 2022) end-to-end document linking, which seeds an artist-constrained artwork pool through the Wikidata creator property (`wdt:P170`); BLINK and GENRE, included as complementary retrievers for explicitly titled mentions; and a single dense bi-encoder that targets implicit mentions by scoring an LLM-rewritten structured representation of the mention against a structured Wikidata view of each candidate. The bi-encoder is queried over two complementary indices in turn – a *restricted view* built from the artist-constrained pool and a *global view* covering every artwork-typed Wikidata item – so that the artist prior is exploited when it suffices and the global index serves as a recall fallback when the work's creator differs from the biographed artist or when ReFinED fails to link the artist upstream. The merged top-k candidates from all four retrievers are then handed to the LLM ranker described in Section 7.4.9.4, which selects a Wikidata identifier or emits `NIL` when no candidate fits. Each mention is linked independently; consistency across mentions of the same entity is enforced indirectly when their per-mention decisions converge on the same Wikidata identifier.

7.4.9.1 Prompt engineering for entity linking

EL operates as a re-ranking task, where for each recognized mention, the dense-retrieval stage produces a ranked list of Wikidata candidates, and the LLM selects the best match or returns `NIL` for a detected OOKB entity. As with the modular structure of our entity recognition prompts, the linking prompts follow the template infrastructure described in Section 7.4.8.3, allowing the role, task, and demonstration layers to be varied independently of the structural template.

The baseline linking variant presents six few-shot examples. Five examples focus on early Michelangelo sculptures, including explicit variants and implicit references. These are evaluated against a unified candidate pool comprising four Michelangelo works, an Arnolfo di Cambio statue, and a Brunelleschi statue. The sixth example addresses `NIL` detection for a generic mention, using a separate candidate pool of three works. This design isolates explicit title matching, close-distractor disambiguation among co-located works, implicit-descriptive matching based on location and material, and generic-mention `NIL` detection as distinct challenges. The model is evaluated solely on its selection of a candidate index or `NIL`.

Few-shot example design. In the recognition stage, we defined few-shot examples from the GT; this means that paragraphs that appear verbatim in an example will be excluded from

evaluation. Correspondingly, the few-shot examples in the EL stage consist of manually selected paragraphs from the Michelangelo biography and a generic-mention case from a Botticelli artwork.

The six examples are selected to cover these scenarios:

- **Explicit title matching:** the *Saint Petronius* and *Rondanini Pietà* examples from Michelangelo, each with the canonical title appearing in the context;
- **Double function of *Angel* for disambiguation:** the *Angel* example is actually an explicit mention (that is its title) but look like an implicit one; therefore, this example is disambiguated with a candidate pool that contains several Michelangelo sculptures co-located at the Basilica of San Domenico (Saint Petronius, Saint Proculus, Angel), and the context's reference to a candelabrum-bearing angel must select the right one;
- **Implicit matching:** the implicit mentions *the statue* and *the marble figure* functions as few-shot examples in which the model must reason from location (Church of S. Petronio façade) and material context to identify the referent;
- **Generic mentions and `NIL` detection:** the generic mention *many works in fresco* from Botticelli's biography works as an example where the surface form matches no specific recoverable artwork in the candidate pool, and the correct answer is `NIL`.

Selecting six examples establishes the minimal set necessary to instantiate all discrimination patterns targeted by the few-shot demonstrations. Examples 1–5 share a fixed six-candidate pool spanning multiple artists and periods, so the model must actively discriminate rather than retrieve by creator alone; example 6 uses a separate three-candidate Botticelli pool to exercise the `NIL` pathway in isolation. Using six examples is large enough to expose the model to multiple within-artist distractors (four Michelangelo works) and at least one out-of-artist distractor pair (Arnolfo di Cambio, Brunelleschi) per demonstration. The `NIL` demonstration uses a smaller three-candidate Botticelli pool because the discrimination it teaches – generic mention $\rightarrow$ `NIL` – does not require multiple distractors. This design rationale follows the well-established observation that the representativeness and diversity of in-context examples drive few-shot performance more strongly than their raw counts (Liu et al. 2021), and that example selection dominates few-shot accuracy in extraction settings (Polat et al. 2025). The pool size is therefore a design consequence of the taught discrimination structure rather than a tuned hyperparameter.

7.4.9.2 Knowledge base provenance

The external KB used throughout the EL pipeline is Wikidata (Vrandečić and Krötzsch 2014), used for both candidate generation and evaluation against the GT. The explicit entity linkers use

versions of Wikidata as existing at the time of their creation. For the retrieval methods focusing on implicit entity linking, all Wikidata queries and precomputed candidate indices are derived from a full Wikidata snapshot from mid 2025. The snapshot is used in the 'global view' setting. In the *artist-constrained* setting, we use a SPARQL endpoint of a 2025 version of Wikidata, hosted on local servers. SPARQL queries are issued against this endpoint so that retrieval results remain deterministic and independent of any post-snapshot edits to Wikidata.

7.4.9.3 Candidate generation query rewriting

An LLM-based query-rewriting stage transforms a mention's surface form and surrounding paragraph context into a structured metadata record that the bi-encoder then uses as the retrieval query in place of the raw mention-in-context string. The metadata-extraction prompt uses two contrastive examples, an explicit mention (Rembrandt's *Night Watch*) in which all fields can be populated, and an implicit mention (a Vermeer interior scene described without a title) in which several fields remain empty. The motivation is representational alignment. The bi-encoder is trained on passage vectors that are themselves structured Wikidata property records (Table 7.13), so a structured query closes the representational gap between the query and the passage. Structuring both sides has the added effect of making the fields look the same regardless of whether the candidate record is a modern Wikipedia summary or a sixteenth-century biography, thereby improving generalizability across different textual sources. This intuition is borne out empirically. We tested this with an input-representation ablation (Section 9.4.4), which shows that querying the index with the LLM-generated structured representation consistently outperforms querying with the raw mention embedded in its paragraph context, across both retrievers and across all mention types. Accordingly, the default ObliquER pipeline uses the structured query.

Candidate generation relies on dense passage retrieval to produce a set of potential candidates. Specifically, an entity mention and its surrounding context are concatenated, then encoded to compute a query vector representing the mention. This process is applied to each entity identified in the entity recognition step (see Section 7.4.8). A concatenation step summarizes the label, description, and several additional properties into passage vectors. The query vector is then compared against all available passage vectors using cosine similarity, efficiently implemented via approximate nearest-neighbor search (Indyk and Motwani 1998).

As a concrete example, consider the following passage from the biography of Tiziano Vecellio:

> [. . .] He made known to Vasari Messer Gian Maria Verdezotti, a young Venetian gentleman full of talent, a friend of Tiziano and passing able in drawing and painting, as he showed in some landscapes of great beauty drawn by him. This man has, by the hand of Tiziano, whom he loves and cherishes as a father, two figures painted

> in oils within two niches, an Apollo and a [IMPLICIT MENTION]Diana[IMPLICIT MENTION].

From the implicit mention *Diana*, the recognition stage generates the following structured metadata:

> **label:** Diana
> **description:** An oil-painted figure of the goddess Diana, placed within a niche, created as part of a pair with an Apollo figure.
> **creators:** Tiziano Vecellio
> **materials:** oil
> **depictions:** Diana
> **instance of:** painting

This structured representation is concatenated into a single passage string. On the Wikidata side, each candidate artwork is encoded into a structured Wikidata record (using the properties in Table 7.13). The bi-encoder is trained on the combined ArtPedia and SemArt training corpus, fine-tuned to score against the structured Wikidata record. The single dense bi-encoder scores the mention representation against all passage vectors in the artist-constrained as well as the global index via cosine similarity, and the top-k most similar candidates from each are merged into the candidate set for subsequent ranking.

7.4.9.4 Candidate generation and ranking

The core components of the ObliquER EL stage are candidate generation and subsequent candidate ranking. Both are described in the following.

Candidate generation in ObliquER. Candidate generation in ObliquER follows the retrieve-then-rank paradigm (Wu et al. 2020, Logeswaran et al. 2019), combining multiple complementary retrievers to assemble a diverse candidate pool for each mention.

As a first step, we run ReFinED end-to-end over the document to link explicit artists (artists mentioned in biographies). The structural prior we exploit here is straightforward. In an artist biography, the biographed individual is typically the creator of the artworks discussed, so for each person ReFinED links we retrieve all artworks attributed to that person in Wikidata via the creator property (`wdt:P170`), yielding a small, high-precision artist-constrained artwork pool. On top of this pool, three further retrievers operate in parallel:

- BLINK and GENRE for explicit mentions. Both models were trained predominantly on named entities and do not generalize well to implicit surface forms, so they serve primarily as a recall backstop for canonically titled works. We include both rather than either alone because their decoding strategies are complementary: BLINK retrieves discriminatively over a fixed entity index, GENRE generates entity names autoregressively with beam search, and combining them increases candidate coverage in an ensemble-like fashion.
- The single dense ObliquER bi-encoder, fine-tuned from `Snowflake/snowflake-arctic-embed-m-v2.0` on the combined ArtPedia and SemArt training corpus (see Section 7.4) and specialized for the structured Wikidata passage view. This targets implicit mentions by scoring a structured representation of the mention against a structured Wikidata property view for each candidate artwork. The same bi-encoder is queried over two complementary indices, the artist-constrained artwork pool (the 'restricted view') and a broad index covering all artwork-typed items in the Wikidata snapshot (the 'global view').

The artist-constrained pool has limited recall when the artwork's creator differs from the biographed artist or when ReFinED fails to link the artist; the global pool serves as the recall fallback in those cases, while the two pools share the same bi-encoder, so the same similarity geometry applies on either side.

The bi-encoder takes a structured representation of the mention generated during the recognition stage (see Section 7.4.9.3 for the structured-representation rewrite) and compares it against a structured Wikidata representation of each candidate artwork, composed from a curated set of properties (see Table 7.13). A prioritization is on the encoding of contextual information most commonly available for implicit artwork mentions in Vasari, such as creator (`wdt:P170`), commissioner (`wdt:P88`), location (`wdt:P276`), iconographic subject (`wdt:P180`), and material (`wdt:P186`).

The retriever combination and per-retriever candidate budget were determined by a greedy selection procedure on the ArtPedia validation set, using ArtPedia rather than the Viewsari dataset to avoid reducing our already small evaluation corpus from the GT. The search first identified the per-retriever budget from $\{4, 8, 16\}$ that maximized candidate recall. Retrievers were then selected greedily by end-to-end F1 under the reranker, so that the single best retriever was added first, followed by each remaining one only if it improved performance. This was done because the order and size of the candidate set can significantly impact the ranker's performance, as we also observed in the greedy selection procedure.

For each entity mention, the surface form is marked with explicit delimiters and concatenated with up to 100 tokens of left context and 100 tokens of right context from the source paragraph to form the query string. This also serves for the linking step in Section 7.4.9.4. The passage

Table 7.13: Wikidata properties used for artwork entity representations in ObliquER.

Wikidata property (IRI)	Label
`rdfs:label`	Label
`schema:description`	Description
`wdt:P571`	Inception
`wdt:P180`	Depicts
`wdt:P195`	Collection
`wdt:P186`	Made from material
`wdt:P88`	Commissioned by
`wdt:P276`	Location
`wdt:P135`	Movement
`wdt:P31`	Instance of
`wdt:P170`	Creator

representation for each Wikidata candidate is a structured plain-text record assembled from the properties listed in Table 7.13: title, description, creators, inception, depictions, collections, materials, commissioner, location, movement, and instance type. Fields absent from Wikidata are omitted rather than left blank, keeping passage representations compact. All embeddings are ℓ_2-normalized before indexing, so that inner-product search is equivalent to cosine similarity.

In the unconstrained setting, the passage vectors are precomputed and stored in a local approximate nearest-neighbor index. In the artist-constrained setting, the SPARQL query is issued at runtime for each biography; the resulting candidate set is then embedded, and exact dot-product similarity is used because the constrained sets are small. Results are cached per artist identifier within a pipeline run to avoid redundant queries.

The artist-constrained index based on `wdt:P170` is cost-efficient because it enables the creation of subsets, such as Michelangelo's or Pontormo's artworks, thereby reducing the number of candidates to retrieve. To retrieve the set of candidate entities from the `wdt:P170`-constrained subset of Wikidata, we issue a SPARQL query, then encode each retrieved artwork as previously described.

To illustrate how implicit entity mentions remain challenging because many artworks lack fields like creator or location, we showcase the explicit artwork mention of *the Nile*, which appears in the following passage from the Michelangelo biography:

> [. . .] And there, to enrich that part, he caused to be placed on certain bases the two ancient figures in marble of recumbent River Gods, each of nine braccia, and of rare

> workmanship, one of which is the Tiber and the other [EXPLICIT MENTION]the Nile[EXPLICIT MENTION]; and between them, in a niche, is to go a Jove. [. . .]

The referent is the sculptural figure of the Nile, corresponding to the Wikidata entity Q113009055. This entity, however, lacks the wdt:P170 creator property in Wikidata, so it does not appear in the artist-constrained candidate set generated for the Michelangelo biography. It cannot be retrieved via the standard pipeline. Instead, in the best-case scenario, the potential candidate for this entity is retrieved with the broader Wikidata artwork subset (from the 'global view' as a backdrop).

In our approach, the OOKB detection performance reported in Chapter 9 measures the model's capacity to identify when no suitable candidate entity is present. This outcome may indicate a genuine NIL case, which requires assignment of a dedicated internal Viewsari IRI. Alternatively, it may reflect insufficient information to resolve the entity, either because the candidate is absent from the retrieved set or because the Wikidata subset lacks the necessary data for identification. These scenarios highlight current limitations and motivate future work on refining candidate generation.

Candidate ranking in ObliquER. For candidate ranking, we employ an LLM as the decision-maker; the prompt structure and few-shot design are detailed below. An LLM is well-suited to this step because it can reason jointly over the mention's surrounding context and each candidate's properties, natively handles OOKB rejection by emitting a NIL label rather than forcing a link to any candidate, and adapts to new domains through few-shot prompting without retraining.

At inference time, each entity mention is presented to the ranker together with its surrounding context, a ranking prompt, and the merged candidate set produced by the four tuned retrievers selected through the greedy procedure of the previous paragraph. The per-retriever budgets retained by that procedure are **16** candidates from the bi-encoder over the artist-constrained pool, **16** from the bi-encoder over the global Wikidata index, **16** from GENRE, and **4** from BLINK, for a combined ceiling of **52** candidates per mention before deduplication. ReFinED does not contribute candidates at this stage; it operates upstream, seeding the artist-constrained pool via wdt:P170 (see the document-level linking step described above). In practice, the four retrievers' candidate sets overlap, so the deduplicated set actually shown to the ranker is typically smaller than 52. This runtime budget is distinct from the in-prompt teaching pool in Section 7.4.9.1, which sizes the candidate lists shown alongside the few-shot demonstrations rather than the candidate set seen at inference time. The ranker returns the best matching candidate or NIL if none fit, indicating an OOKB entity. The ranking prompt uses role-based prompting and includes few-shot examples for explicit and implicit mentions, as well as the OOKB case. Figure 7.29 sketches a representative ranker call, in which a mention with its paragraph context is presented to the ranker

alongside the top-k retrieved candidates, and the ranker emits either the Wikidata identifier of the chosen candidate or a Viewsari-internal OOKB IRI.

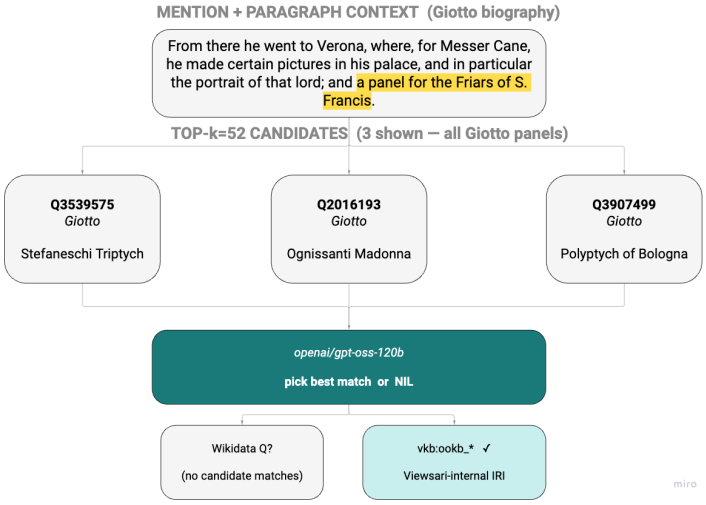


Figure 7.29: EL ranker as an LLM decision-maker, applied to the Vasari implicit mention *"a panel for the Friars of S. Francis"* from the Giotto biography. The relevant passage describes Giotto's commissions in Verona for *Messer Cane* (Cangrande della Scala), and the highlighted mention refers to a panel painted there for the Veronese Franciscans. Together with up to 52 retrieved candidates (three plausible Giotto-panel distractors are shown here for compactness – the *Stefaneschi Triptych*, the *Ognissanti Madonna*, and the *Polyptych of Bologna*, none of which is the correct referent), the mention and its paragraph context are presented to `openai/gpt-oss-120b`. Because the specific Veronese Franciscan panel is not individually cataloged in Wikidata, the ranker's correct decision in this case is to emit the Viewsari-internal OOKB IRI `vkb:ookb_*` rather than to force a link to any retrieved candidate. The pipeline preserves OOKB status rather than forcing a wrong link.

The ranking prompt is organized as a two-part structure, namely a system message that frames the model as an expert EL assistant instructed to return only a candidate index or `NIL`, with no additional output; and a user-turn template that renders the task description, the candidate list (a numbered sequence of plain-text records using the same field serialization as the passage representations), and the mention with surrounding context. The template includes the six fixed few-shot examples from Section 7.4.9.1.

7.4.10 Extending the pipeline to a contemporary art-historical corpus: ArtPedia

To assess the transferability of the ObliquER pipeline beyond sixteenth-century biographical prose (even when translated into 20th-century English), and to provide an extended evaluation in Chapter 9, we apply the recognition and linking stages to ArtPedia.

ArtPedia is an artwork dataset of 2,930 paintings, each linked to sentences from Wikipedia articles. These sentences are labeled as either visual sentences, which describe the artwork's pictorial content, or contextual sentences, which cover historical background, authorship, artistic influence, and exhibition location. The dataset contains 28,212 sentences in total, of which 9,173 are visual and 19,039 are contextual. For ObliquER, ArtPedia offers a modern encyclopedia perspective rather than Renaissance biographies, and the artworks are the main focus of the articles rather than being mentioned only briefly.

This dataset serves as an addition to ObliquER's strategy on the Viewsari dataset because, while it is functionally similar, it lacks a clear separation between implicit and explicit artwork mentions; instead, a combination of visual and contextual sentences creates a *synthetic* implicit environment. References are implicit because the sentences focus primarily on content description and avoid the title (some sentences do include canonical titles; those are marked as explicit mentions). Accordingly, we can simulate the environment of Viewsari by creating a counterpart in a modern, quantitatively larger setting. Essentially, this makes ArtPedia a benchmark for comparing ObliquER's performance in entity recognition and linking for implicit entity mentions (see Section 9.4.7). The ArtPedia dataset functions as a silver-standard GT.

Originally, in ArtPedia, an artwork is listed as shown in Figure 7.30. We map this to the following structure in the data preparation (Section 7.4.10.1):

Wikipedia dump extraction. Since ArtPedia's sentences were originally drawn from Wikipedia, we rebuild the full article texts using a 2019 Wikipedia dump to match the dataset's temporal scope. The dump is retrieved from the Internet Archive[25] and each record carries the article identifier, URL, title, and the full text.

[25] Specifically, the `enwiki-20190120-pages-articles.xml.bz2` dump, archived at the Internet Archive: `https://archive.org/download/enwiki-20190120/enwiki-20190120-pages-articles.xml.bz2`. The download and processing steps are documented in the project repository.

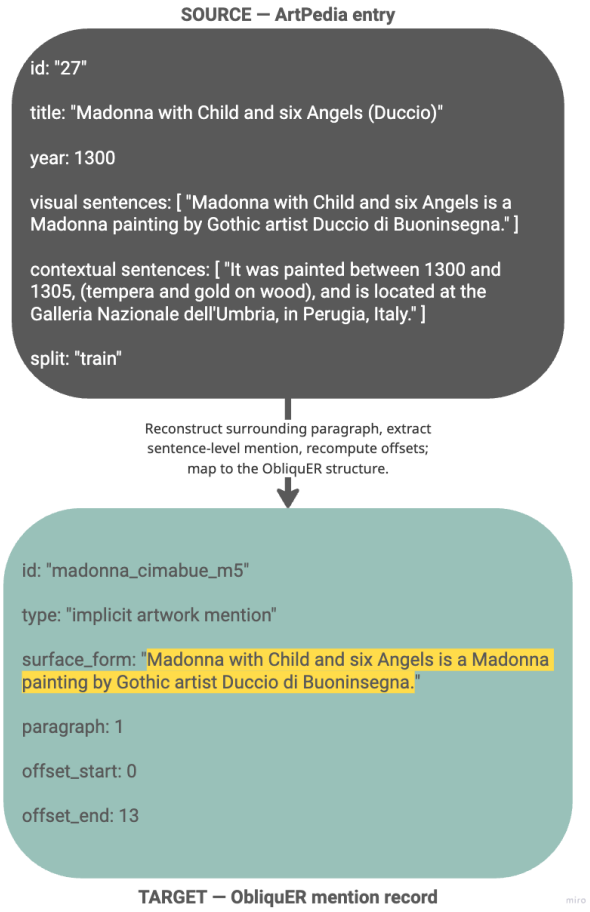


Figure 7.30: Mapping the ArtPedia entry of Madonna (top) to the representation of an ObliquER paragraph-local mention record (bottom). The highlighted surface form is the sentence-level mention (either visual or contextual in the original), selected from the reconstructed paragraph; `type` carries the Viewsari mention type.

7.4.10.1 Dataset preparation and ground truth construction

We used the artwork titles (as shown for the Madonna example in Figure 7.30) from the ArtPedia dataset to find *explicit artwork mentions* by matching them with strings in the visual and contextual sentences. When a sentence is found, the mention type is determined by the rule that if the base title of the artwork – the portion of the Wikipedia title before any parenthetical disambiguation, e.g. *Maestà* from *Maestà (Cimabue)* – appears within the sentence, the mention is typed as *explicit* and the surface form is set to the base title; otherwise, the mention is typed as *implicit* and the full sentence is taken as the surface form. Character offsets are computed within the paragraph text, and each mention carries the Wikidata identifier of the corresponding artwork. The necessary alignment links the ArtPedia title to a Wikidata identifier via the Wikipedia and Wikidata APIs. Since ArtPedia uses Wikipedia as its source, annotated artworks already exist in Wikidata, resulting in a low OOKB entity ratio.

7.4.10.2 Prompt development

The ArtPedia prompts follow the same modular template structure as the Viewsari prompts (see Section 7.4.8.3). For each Wikipedia article, only those paragraphs containing at least one annotated mention from the ArtPedia GT are passed through the pipeline, reducing inference costs without affecting the evaluation, and the prompts are pre-rendered in one file per paragraph per article.

The main differences between the prompt templates for ArtPedia and Viewsari are the role descriptions and the few-shot examples provided. Instead of examples from Vasari's text, we use examples from contemporary Wikipedia articles. These static examples include a paragraph about *The Scream* by Edvard Munch to show explicit and implicit mention annotation in an encyclopedic style:

> The Scream is the popular name given to a composition created by Norwegian Expressionist artist Edvard Munch in 1893. The original German title given by Munch to his work was (The Scream of Nature), and the Norwegian title is (Shriek). The agonized face in the painting has become one of the most iconic images of art, seen as symbolizing the anxiety of modern man.

From this paragraph, we extract two mention records (Figure 7.31): one *explicit* mention whose surface form is the canonical title *The Scream* and whose in-text span begins at the very start of the paragraph, and one *implicit* mention whose surface form is the third sentence's descriptive paraphrase of the painting. Both records share the same Wikidata identifier (`Q471379`), and

together they illustrate the two ObliquER mention types instantiated within a single Wikipedia paragraph.

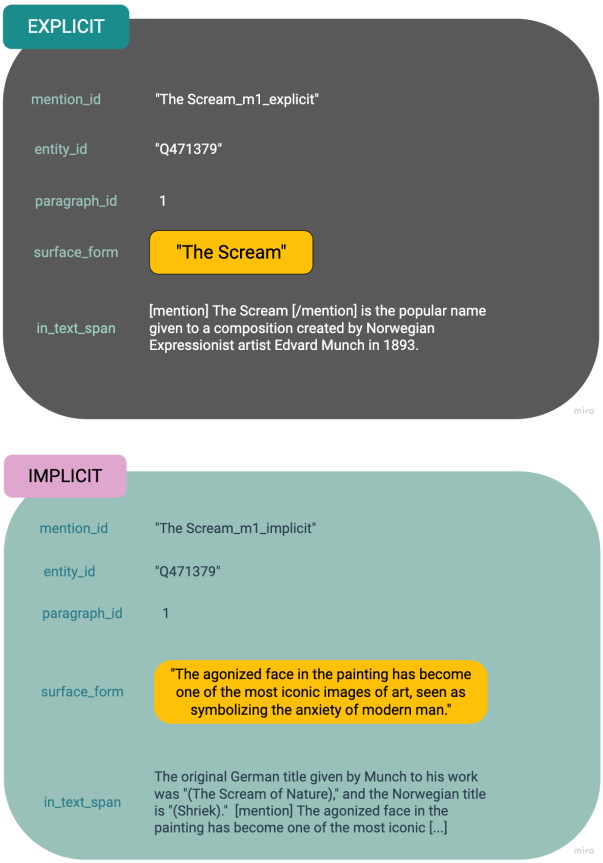


Figure 7.31: The two ObliquER mention records produced from a single ArtPedia paragraph about *The Scream*. **Top – explicit mention:** the canonical title *The Scream* appears verbatim, the `surface_form` is short, and the in-text span anchors at the start of the paragraph. **Bottom – implicit mention:** the third sentence describes the painting without naming it, the `surface_form` is the full descriptive sentence, and the in-text span carries the preceding context. The shared `entity_id` `Q471379` links the two mentions to the same Wikidata entity. The in-text-span tags are abbreviated to `[mention]` / `[/mention]` for display.

ArtPedia keeps the pilot's empirical surface within the same domain (art-historical artworks linked to Wikidata) as Viewsari, exercising the ObliquER pipeline on a second corpus with a different period and register.

A second pilot study moves the pipeline out of art history entirely – onto a 19th-century German Masonic chronicle, with persons, organizations, and locations as the target entity types, and FactGrid as the primary KB – and is treated in full in Chapter 10 (see Section 10.2.1).

7.5 Knowledge graph population

The Viewsari ontology developed in Section 7.3, in the Phase II form described in Section 7.3.2 and published as a single Turtle file in the project repository,[26] functions as the schema for the populated KG. Every individual instantiated by the population pipeline is typed following the class hierarchy and properties from the ontology, and every provenance link follows the activity taxonomy refined in Section 7.3.2.2. The remainder of this section describes how we build the graph against that schema, in the order in which the underlying data were originally obtained, and the resulting Turtle serialization is versioned alongside the ontology in the same repository.[27] The population pipeline is positioned alongside the digital-humanities KGs surveyed in Section 5.2.3 – in particular, the Bisticci-letters HiCO graph (Section 5.2.3.1), the Polifonia KG (Section 5.2.3.2), the European Olfactory KG (Section 5.2.3.3), and FactGrid (Section 5.2.3.6).

7.5.1 Bibliographic information

The pipeline takes as given a bibliographic and structural backbone that mirrors the ontology's FaBiO levels (Section 7.3.1.2), instantiating in concrete RDF the work/expression/manifestation/item stratification surveyed in Section 5.2.2.3. Each of the ten volumes of the 1912 de Vere translation, each of the 165 biography instances it contains, and every page within each biography is materialized as an expression–manifestation pair, the manifestation linking out to the corresponding Project Gutenberg web versions; page ranges follow the printed page numbers embedded as anchors in the Gutenberg source, and the end page of a biography is fixed at the page immediately preceding the start page of the next biography in the same volume. Paragraphs, which were previously extracted from the same source, are merged in page order into a unified text string, and each merged paragraph is anchored to the page-manifestation nodes of its start and end pages (cf. Table 6.3). On top of this, four content layers are added in turn – persons and their co-occurrences, the manually annotated artwork GT, and the ObliquER recognition outputs – and a small auxiliary layer of facsimile data is produced for the web interface but deliberately omitted

[26] Viewsari ontology serialization: `https://github.com/ISE-FIZKarlsruhe/viewsari/blob/main/data/ontology/viewsari_ontology.rdf`.

[27] All entities reside under the namespace `vkb:` = `https://viewsari.ise.fiz-karlsruhe.de/kb/1.0#`.

from the graph itself. Aggregate population statistics for every layer are reported in Section 8.3; this section describes only how the layers are constructed. This population has already been visually illustrated by Figure 7.19.

7.5.2 Persons

The person layer is derived from the manually compiled Index of Names that concludes each of the ten volumes of the de Vere translation (cf. Section 7.2; the use of a curated, source-internal index as the spine of a person layer mirrors the methodology of the Bisticci-letters HiCO KG surveyed in Section 5.2.3.1, which similarly anchors person identification in a manually compiled source-resident apparatus before reconciliation against external authorities. For each volume, the extraction step reads the individual mentions of named entities, along with their provenance information, for in-paragraph anchoring; these per-volume mentions are then deduplicated across volumes by name/coreferents to yield one person per name, with alternative labels carried over from the source. A reference list of pre-1568 Wikidata persons reconciles artists with Wikidata identifiers using fuzzy label matching; unmatched persons keep their Viewsari-internal IRIs.[28] Every mention is materialized as an annotation linked it to a `doco:TextChunk` of the source paragraph through an `oa:TextPositionSelector` that encodes the exact character span, and each annotation is provenance-linked to the `viewsari:named entity recognition` activity that produced it; a separate `viewsari:coreference reconciliation` activity is recorded for the merging step that yields the deduplicated person individuals. Figure 7.32 shows this example for Michelangelo in the KG.

7.5.3 Co-occurrences

Each co-occurrence (cf. Section 7.2.1) is materialized as a first-class Viewsari individual, following the pattern fixed in Section 7.3.2.2; it carries two links to the persons involved, an in-paragraph link to the source paragraph, and a `prov:wasGeneratedBy` link to a `viewsari:co-occurrence analysis` activity instance. The corresponding paragraph-level annotation layer mirrors the structure used for person mentions, so that mentions, co-occurrences, and – in the second stage – artwork mentions all share the same provenance-anchoring pattern and can be reached by SPARQL queries that need not know an annotation's origin.

[28] This is an artifact from the previous project run by Santini et al. (2022). They used this file to reconcile artist mention, too: `https://github.com/ISE-FIZKarlsruhe/vasari_network/blob/main/linking_vasari_wikidata/artists_before_1568.csv`.

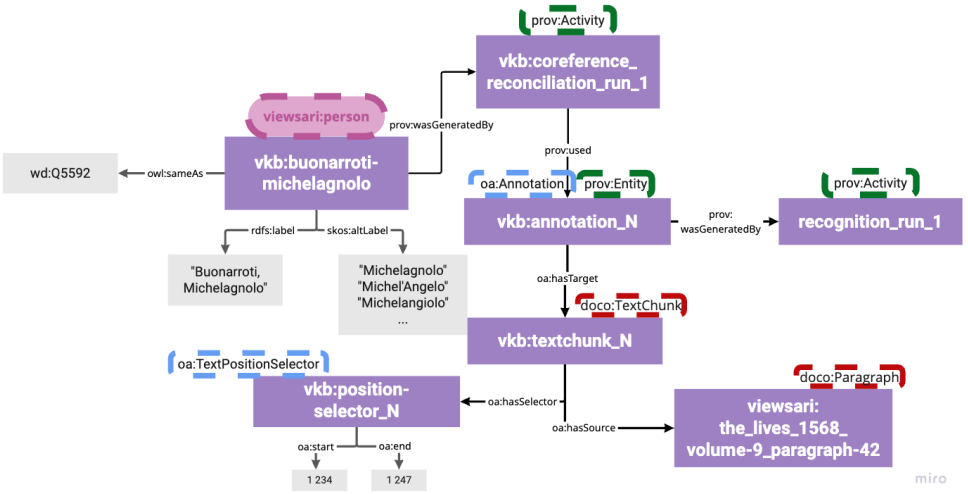


Figure 7.32: Example person node in the populated Viewsari KG, centered on `vkb:buonarroti-michelagnolo` (typed `viewsari:person`). Outgoing edges show the canonical label, the `owl:sameAs` link to `wd:Q5592`, three representative `skos:altLabel` variants harvested from Wikidata (with an ellipsis standing in for the remaining alternative spellings), and the `prov:wasGeneratedBy` link to the `vkb:coreference_reconciliation_run_1` activity. A second branch shows one mention annotation: the `oa:Annotation` resolves through `oa:hasTarget` to a `doco:TextChunk`, whose `oa:hasSelector` carries an `oa:TextPositionSelector` with explicit `oa:start`/`oa:end` character offsets, and whose `oa:hasSource` resolves back to the originating `viewsari:paragraph` instance. The annotation itself is provenance-linked to the `vkb:named_entity_recognition_run_1` activity. Literal values are drawn as rounded rectangles, IRI-typed individuals as ellipses, and external IRIs as dashed ellipses.

7.5.4 Ground-truth mentions and entities

For the manually annotated GT corpus produced as described in Section 7.4.4, each annotation record carries a mention identifier, a mention type – one of *explicit artwork mention*, *implicit artwork mention*, *coreferent*, or *generic mention* – the surface form, character offsets, an entity-cluster identifier shared by all mentions of the same artwork, a Wikidata identifier where one was found at annotation time, an OOKB flag where none was, a `refers_to` link for coreferent mentions, and an optional Web Gallery of Art identifier.

The alignment between this annotation schema and the Viewsari ontology, summarized in Table 7.14, shows that the mention identifier becomes the URI of a Viewsari mention instance; the mention type defines that the entity instantiates as one of the four mention subclasses; the surface form becomes the textual body of the annotation; the character offsets become the bounds of the position selector; the entity-cluster identifier becomes the URI of the corresponding artwork node; the Wikidata identifier links the artwork to its external counterpart via `owl:sameAs`; the

coreference reference points to the antecedent mention; and the optional Web Gallery of Art identifier becomes an additional `rdfs:seeAlso` link. The same schema is reused for the ObliquER outputs in Section 7.5.5, so that manual and automated mentions are structurally identical in the graph and can be compared using SPARQL queries.

Table 7.14: Mapping from annotation record fields to Viewsari ontology constructs. The same schema is shared by the manually annotated GT and by the ObliquER recognition outputs of Section 7.5.5.

Annotation field	**Ontology construct**
mention identifier	URI of the mention instance
mention type	mention subclass (4-way)
surface form	annotation body literal
start / end offset	start / end of the position selector
entity cluster id	URI of the corresponding artwork instance
Wikidata identifier	`owl:sameAs` link to Wikidata
coreference target	link to antecedent mention
Web Gallery of Art id	`rdfs:seeAlso` link to the Web Gallery of Art

GT ingestion is split into two provenance activities that mirror the conceptual separation between recognition and linking. A `viewsari:named entity recognition` activity summarizes every annotated paragraph and produces all GT mentions, and a `viewsari:entity linking` activity produces the corresponding artwork entities. Both activities are associated with the human annotator as a `prov:Person`, dated to the annotation process.

Entity identity is fixed through mentions linked to Wikidata, which produce an artwork node carrying an `owl:sameAs` link to the Wikidata IRI; mentions with an OOKB IRI produce a Viewsari-internal artwork node; and mentions that carry neither do not generate an artwork entity. Each artwork node is provenance-linked to all contributing mentions via `prov:wasDerivedFrom`, and same-identifier mentions across different biographies collapse into a single artwork entity, thereby encoding cross-biography co-reference at the entity level.

A side product of the GT annotation process – digitized facsimile pages of the 1912 edition, with their word-level bounding boxes – is used exclusively for the side-by-side facsimile overlay of the web interface, and is therefore described together with the application that consumes it (Section 7.7.2). Neither the page images nor the digitized facsimile enter the KG, and the ontology imposes no class for them; they are mentioned here only so that the reader can locate them in the repository alongside the other GT artifacts.

7.5.5 ObliquER output integration

The second population stage integrates the outputs of the ObliquER pipeline into the same graph through the same schema mapping (Table 7.14). The canonical content layer reported as part of the deployed KG is produced by the re-authored *ontology-guided* recognition variant at $k = 2$ introduced for population Run 2 (Section 7.4.8.4, Section 8.3.5). The few-shot variant and the original Run 1 ontology-guided variant are ingested alongside it as parallel mention layers, while the CoT variant is retained in the repository but not loaded into the deployed graph, because the few-shot and ontology-guided variants perform most promisingly (see Section 9.4). This serves as a parallel to the GT ingestions, so that both manual and computational extraction/annotation processes on Vasari's text can be followed in detail with the full Viewsari provenance trail (see Figure 7.33).

For the linking stage, the deployed entity-link layer is produced exclusively by the few-shot ranker variant of Section 7.4.9.1, operating over the re-authored ontology-guided recognition layer in a single linking run.

The ingestion step reads the ObliquER output and resolves character offsets post hoc by aligning the bracketed surface form in the excerpt with the paragraph text, falling back to a direct substring search when the excerpt cannot be found.

Provenance is recorded at the granularity of one prompt response per paragraph, operationalizing in concrete graph form the proposal of Section 5.1.3.3 that prompt strategies are themselves provenance objects to be materialized in the KG rather than treated as ephemeral configuration, and instantiating the PROV-O activity/agent/entity triad surveyed in Section 5.2.2.2 at the granularity of individual extraction runs. An umbrella `viewsari:named entity recognition` activity is created for the prompt strategy as a whole (it follows the design described in Section 7.4.8.3) and associated with the LLM agent, timestamped in the inference logs.

Under that umbrella, each per-paragraph extraction is recorded as its own `viewsari:named entity recognition` sub-activity, linked through `prov:wasInformedBy` to the umbrella, through `prov:used` to both the target paragraph and – where the sliding-window context of Section 7.4.8.1 introduces one – the preceding context paragraph, and through a further `prov:used` link to a `prov:Entity` representing the used prompt, carrying its full text for inspection. Each mention produced by the model is materialized both as an `oa:Annotation` and as a `prov:Entity`, typed with the appropriate mention subclass, with its surface form attached as an `oa:TextualBody` and its target chain anchoring it to a `doco:TextChunk` of the source paragraph through an `oa:TextPositionSelector`; the `prov:wasGeneratedBy` link points to the enclosing per-paragraph sub-activity, so that the chain from any extracted mention back through its prompt, its paragraph, and the model run is recoverable in a single SPARQL hop. Figure 7.33

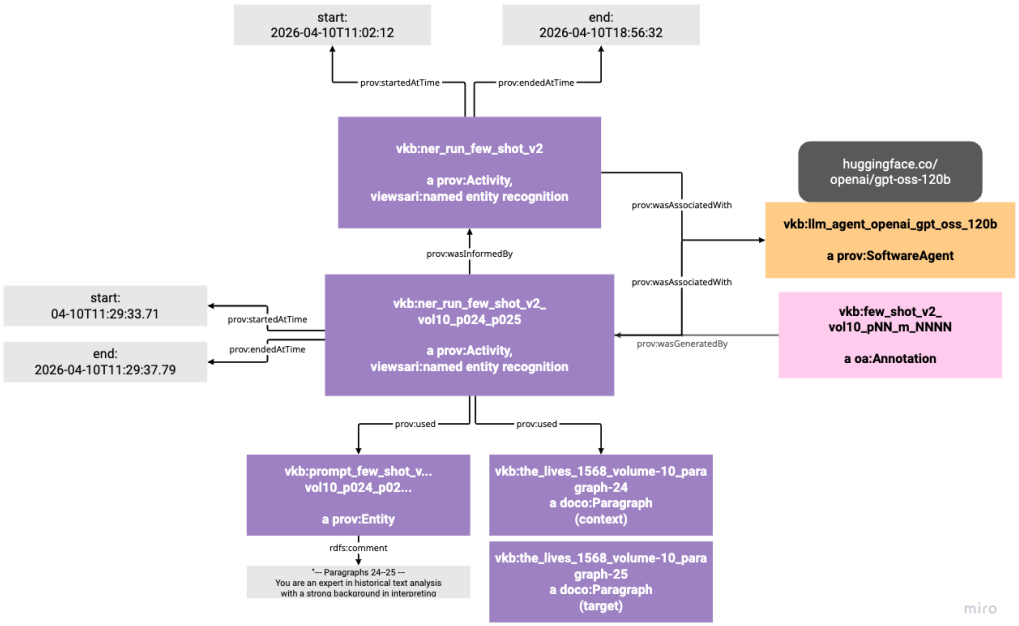


Figure 7.33: Provenance shape of an ObliquER per-paragraph NER sub-activity in the populated KG, instantiated on `vkb:ner_run_few_shot_v2_vol10_p024_p025`. The sub-activity (center) is typed both as `prov:Activity` and as `viewsari:named entity recognition`, and carries `prov:startedAtTime`/`prov:endedAtTime` from the inference log. `prov:wasInformedBy` links it upward to the strategy-level umbrella `vkb:ner_run_few_shot_v2`, which itself records the wall-clock span over the whole few-shot run and is associated, like the sub-activity, with the LLM agent `vkb:llm_agent_openai_gpt_oss_120b` (an `rdfs:seeAlso` link points to the model's Hugging Face page). Three `prov:used` edges record what the sub-activity consumed: the rendered prompt as a `prov:Entity` carrying the full prompt text on `rdfs:comment`, the context paragraph (here Vol. 10, paragraph 24), and the target paragraph (Vol. 10, paragraph 25).

shows the resulting activity shape for one concrete sub-activity drawn directly from the serialized graph.

Entity-level reconciliation from the ObliquER linking stage, described in Section 7.4.9, adds identifiers to the entities, which are then ingested into the KG.

Entity linking is applied only to this canonical recognition layer. Mentions of the re-authored ontology-guided variant are clustered into artwork entities and reconciled with external identifiers, whereas the few-shot and the original ontology-guided recognition layers remain in the graph as fully provenanced `oa:Annotation` nodes that are not promoted to the entity layer. Retaining these superseded runs rather than overwriting them follows directly from the accountability commitment that motivates the ontology (Section 11.1); an earlier interpretive run stays queryable alongside the current one instead of being discarded, and the web interface renders the variants side by side

for the same artwork in its explorer view (Section 7.7.2). The resulting entity counts, together with the cross-variant reconciliation, are reported in Section 8.3.5.

The end-to-end consistency of our process chain, as well as the satisfiability of the CQs derived during requirements engineering (Section 7.3.1.1, Section 7.3.2.1), is validated through the ruleset, which is documented together with the deployment configuration in Section 7.6, and the corresponding counts are reported in Section 8.3.

7.6 Knowledge graph deployment and evaluation infrastructure

The serialized graph, as described in the previous section, constitutes the static output of the population pipeline. To render this graph queryable for the evaluation reported in Chapter 9 and to be integrated by the web application in Section 7.7, we host the graph via a SPARQL endpoint configured with a reasoner. In addition, we introduce an evaluation sub-graph that instantiates the competency-question catalog, developed during requirements engineering (Section 7.3.1.1, Section 7.3.2.1) as executable queries against the endpoint. Both the deployed graph and the evaluation layer are versioned within the project repository and together form the core methodological contribution of this dissertation. The deployment configuration specifies the ruleset governing all reported precision, recall, and coverage metrics, while the evaluation view substantiates the claim that the populated graph meets the requirements under the Viewsari ontology.

7.6.1 Containerized deployment with GraphDB

The runtime stack is containerized for three services, namely the Viewsari web application, an Ontotext GraphDB triple store, and an `nginx` reverse proxy.[29][30] The web application image is built from the project source, incorporating the triple store image pulled directly from GraphDB. The web application's templates and static assets are baked into its image at build time, ensuring production deployments are reproducible from a single image hash.

Two non-default settings of the GraphDB repository configuration are decisive for the evaluation in Chapter 9. The ruleset is set to RDFS-Plus optimized, which materializes the closure of subclass, subproperty, inverse-of, transitive-property, and symmetric-property entailments at query time. All precision, recall, and coverage figures reported in Chapter 9 are therefore computed

[29] GraphDB: `https://graphdb.ontotext.com/`
[30] nginx proxy: `https://nginx.org/`.

against the same materialized closure that backs the public website; that closure typically lands at approximately 2.46M triples (2,455,976 in the canonical run reported in Section 8.3) after ingestion of the 2,448,545-triple raw graph.

7.6.2 Materializing rule-based inference for offline evaluation

Five of the Phase I CQs (CQ I.3, CQ I.12, CQ I.19, CQ I.21, CQ I.22) require a direct *has-anchor-in-paragraph* edge between a co-occurrence or person individual and its source paragraph. The population pipeline does not write that edge, and the link from a co-occurrence to its paragraph is instead encoded indirectly through its provenance chain. To make the relevant CQs answerable in one hop, we materialize the closure rules offline as a SPARQL UPDATE script (Harris and Seaborne 2013) and store the result in a separate inferred-graph serialization, available on GitHub.

The first rule[31] ensures that the anchor edge is propagated along the chain, linking artwork entities to their generating annotations. In the canonical build (population Run 2), this rule materializes 5,855 new triples, since the entity-layer artworks clustered in that build do not yet carry the in-paragraph edge from the ingestion process. The second rule interprets the label of each co-occurrence individual – which carries a string of the form *"Vol. N, Para. M"* produced by the population step – and asserts the corresponding anchor edge between the co-occurrence and the paragraph instance with matching identifiers. This step materializes 632 new triples. The third rule extends the second by traversing the involves property, thereby asserting the relevant edge from each participating person to the same paragraph, materializing a further 944 triples, for 7,431 derived triples in total (Section 8.3). When present, the inferred-graph serialization is preferentially loaded by the evaluation runner, superseding the raw graph.

Materializing these rules offline, rather than embedding them in a custom triple-store ruleset, enforces a clear separation between the authoritative triples generated by the population pipeline and the derived triples introduced solely to address specific CQs. This distinction operationalizes the population–evaluation boundary defined in Section 7.5. The raw graph records extracted information, while the materialized inference closure captures what can be systematically derived from that extraction according to a documented ruleset.

[31] Rules are based on OWL 2 RL (Motik et al. 2012)).

7.6.3 Competency-question evaluation

The catalog of CQs produced during requirements engineering is operationalized as an executable evaluation. Every SPARQL-answerable question is encoded as a single query file sharing a common prefix block, and a machine-readable index annotates each question with its requirements cluster, its target persona (Section 7.4.2), its complexity class (*simple*, *multi-hop*, or *aggregation*), an *expected-gap* flag for cases where the query is correctly modeled but its answer requires data that has not yet been ingested (e.g. extracted historical events or PMI scores), and a pointer to the corresponding query file. Questions that cannot be answered by SPARQL alone – FAIR-principle compliance, teaching reuse – are flagged as descriptive and accompanied by a free-text rationale rather than a query, so that the catalog covers all 70 CQs and not only those that admit a SPARQL formulation.

In the evaluation, the populated graph is loaded into an in-memory RDF library. Each query result is classified into one of four outcomes following a rule table that distinguishes *fully-answerable* (an `ASK` returns `true` or a `SELECT` returns at least one row), *partially-answerable* (the query is correctly modeled but returns zero rows because the underlying data has not been ingested, as recorded by the expected-gap flag), *unanswerable* (the query fails to parse or raises a runtime error), and *non-SPARQL* (the question is descriptive). Parameterized slots in the queries – *the biography of artist X*, *co-occurrence* C, *location* L – are pinned to representative individuals in the catalog so that every query is immediately runnable, and substituting alternative individuals requires only an edit of the corresponding query file. Successive evaluation reports are timestamped artifacts of successive ontology iterations and are referenced individually in Chapter 9.

7.6.4 Reasoner-level evaluation with HermiT

The competency-question evaluation establishes that the populated graph supports the queries derived from the requirements catalog. A complementary evaluation is required to establish that the ontology itself – independently of the full population, and at the level of OWL 2 DL semantics – is logically consistent and that its intended class hierarchy is derivable by a description-logic reasoner. This second evaluation is operationalized against a synthetic ABox of 27 hand-crafted individuals (323 triples) that provides at least one direct instance per class and one usage per object property, including a co-occurrence with two distinct participation links chosen to satisfy

the qualified-cardinality restriction on the co-occurrence class. The combined TBox and synthetic ABox are loaded into Protégé and classified with HermiT 1.4.3.[32]

We record the following consistency verdict, comprising the inferred subclass axioms beyond those directly asserted in the TBox; the inferred class assertions on the synthetic individuals – in particular, whether the four mention subclasses are entailed members of the mention class and whether artworks, persons, locations, and organizations are entailed members of the *extracted content* superclass and of the PROV *entity* class; a test of expected entailments such as *Person and not Artwork*, *Cooccurrence that involves at least two extracted contents*, and *Coreferent and refersTo some Mention*; and the wall-clock time for consistency checking, classification, and realization. The *imports* chain to FaBiO and PROV-O is removed prior to reasoning because both upstream ontologies ship SWRL rules with built-in atoms that HermiT does not support; their external superclasses remain in the file as opaque IRIs and have no effect on consistency within the Viewsari namespace.

The reasoner-level evaluation is restricted to the synthetic ABox because the examples show different patterns of the ontology; HermiT's complexity does not scale to the fully populated graph of $\sim$2.45M triples, and a representative reasoning-time evaluation against the full graph would not terminate within practical bounds. The full-graph behavior is therefore covered by the competency-question evaluation rather than by description-logic reasoning, and the two evaluations are reported jointly in Chapter 9 so that one offers a consistency claim, while the other offers a perspective on answerability.

7.6.5 Derived data layer for the web application

The website described in Section 7.7 does not query the live SPARQL endpoint for routing-time interactions, since each page load through the endpoint would impose a measurable round-trip cost, and the same data is needed in slightly different shapes by the entity-detail pages, the knowledge-base browser, and the KG explorer. Instead, three offline build steps produce derived data artifacts that the application loads once at startup and serves as static JSON.

The first build step parses the populated graph and extracts three entity classes plus two mention lists, namely persons with their alternative labels and Wikidata identifiers; artworks with their paragraph anchors, occurrences, and source-category tag (GT, ObliquER, or other); co-occurrences with their participating persons; and the per-strategy lists of explicit and implicit

[32] HermiT 1.4.3 reasoner: `http://www.hermit-reasoner.com/`.

ObliquER mentions for the knowledge-base browser. The second build step produces the pre-computed force-directed subgraphs that back the KG explorer view, one graph per seed entity, together with a top-level manifest. The third build step produces, for each prompt strategy and each biography, a graph whose nodes are biographies, paragraphs, mentions (typed by the four mention-type classes), per-paragraph NER sub-activities, prompt instances, and the LLM agent, and whose edges encode the corresponding ontology properties (anchor-in-paragraph, was-generated-by, used, was-associated-with, and the coreferent link). Each edge additionally carries a reverse label so that the info panel can show, for example, *contains* when a paragraph is selected and *is in paragraph* when a mention is selected. All three build steps are deterministic and idempotent. Re-running them after a population change produces a new set of derived artifacts that the application picks up on the next restart, without requiring any change to the application code.

7.7 Website and user interface

The Viewsari web application constitutes an interactive research environment for exploring the KG detailed above. It serves as a demonstrator, exposing the outputs of the ObliquER pipeline and the ontology-grounded population process (Section 7.5) to art-historical researchers without requiring expertise in SPARQL or RDF. The application ensures that each view explicitly surfaces the ontological commitments of the Viewsari ontology, including bibliographic hierarchy, annotation provenance, and EL coverage, which are presented through the interface, and it becomes visible that the underlying KG does not aim to represent the work as a fact but as a complex system of extraction processes and modeling commitments. Not all components and their development are described in detail; implementation details can be found in the GitHub repository.

The user-facing requirements collected during requirements engineering (Table 7.2, *User-facing functionality* block) constitute the construction baseline for the application, namely index search for extracted resources, semantic and visual search, open-ended exploration without a pre-formed RQ, visibility of connection strength between resources, and visualization of resources and their connectedness. These requirements were elicited from the same domain-expert sessions that produced the ontology requirements but were deliberately kept separate from the ontology catalog, since they constrain the system layer rather than the representational schema. The mapping from each requirement to the deployed interface view that addresses it is reported in Table 8.13 of the Results chapter.

Functions retrieved from the descriptive competency questions. The catalog's *descriptive* (non-SPARQL) CQs – the 18 entries that the ontology evaluation discharges outside

SPARQL coverage (Section 9.1.3) – are answered by this application and its documentation rather than by a graph query. The KG explorer (Section 7.7.4) serves the open-ended exploration questions `CQII.5`, `CQII.30`, and `CQII.33`, to the extent that the underlying entity types are populated; their location-, event-, and style-dependent facets inherit the population gaps reported for the SPARQL-translatable CQs (Section 9.1.3). The SPARQL editor and endpoint (Section 7.7.5) answer the data-access questions `CQI.24` (semantic search), `CQI.26` (endpoint and dump availability), and `CQII.27` (federated querying, through the illustrative `SERVICE` template), and the embedded WIDOCO documentation portal answers the modeling-rationale question `CQII.41`. The remaining reuse- and teaching-oriented questions (`CQI.25`, `CQI.29`, `CQII.28`, `CQII.29`, `CQII.31`, `CQII.32`) are addressed by the repository documentation and licence, while `CQII.10` (artwork collaborative effects across shared locations or styles), `CQII.14` (geographical map), `CQII.17` (cross-disciplinary connections), `CQII.18` (historical-event enrichment), and `CQII.23` (parallel-translation comparison) are deferred to future work pending the population of the relevant entity types.

Design references. The interface design draws on a focused set of references cataloged in Section 5.2.6. The Sampo model (Section 5.2.6.2), with its three canonical view types – landing, perspective, and entity pages – and its facet-driven rather than query-driven entry into the graph, informs the overall information architecture of the application, in particular the relation between the homepage gallery, the knowledge-base browser, and the entity-detail pages (Hyvönen 2023, Ikkala et al. 2022). The InTaVia entity landing pages (Section 5.2.6.2), which integrate an ego-network view with a chronological timeline of an entity's productions, motivate the inclusion of interactive node-link views rather than tabular result sets alone (Windhager et al. 2024). The WebVOWL and Graffoo conventions (Section 5.2.6.1) inform the ontology-documentation portal embedded in the site; the interactive ontology graph (Figure 5.10), implemented with D3.js v7 (Bostock et al. 2011), reuses the color-coded module separation of those traditions (Lohmann et al. 2016, Falco et al. 2014). Underlying these exemplars are two broader strands of literature that the application takes as design foundations: the CH Linked Data publishing and portal-design tradition (Hyvönen 2012) and the complementary *generous-interfaces* approach to digital cultural collections (Whitelaw 2015), which jointly motivate the gallery-first landing page, the multi-class knowledge-base browser (Section 7.7.3), and the explorer's open-ended graph traversal in place of a single query-and-result entry point; and the visualization-design typology of Munzner (2014), whose *what*/*why*/*how* decomposition is used implicitly to distinguish the tabular entity-detail views (high-precision lookup of a known item) from the force-directed exploratory views of the KG explorer (Section 7.7.4, exploratory browsing without a pre-formed RQ). Beyond these design exemplars, implementation references are taken directly from the documentation portals of the underlying tools.

7.7.1 Architecture and technology stack

The server component is implemented in *FastAPI*, an asynchronous Python web framework that provides automatic OpenAPI documentation and dependency injection.[33][34] HTML is rendered server-side; the front end relies on vanilla JavaScript and CSS without external framework dependencies.

The application is one of three services in the containerized stack described in Section 7.6.1. Essentially, it handles all routing and template rendering; the triple store hosts the populated graph via a SPARQL endpoint, with the materialized inference closure described above; and an `nginx` reverse proxy terminates external traffic. The artifacts that the application builds on at startup – the per-biography GT records and the derived data layer described in Section 7.6.5 – are mounted into the container from the host data directory, while the application's templates and static assets are baked into its image at build time, so deployments are reproducible from a single image hash.

At startup, the application loads four data sources into memory:

1. **Annotation records.** Per-biography records produced by the ObliquER pipeline, each containing paragraph text, mention spans with character offsets, entity identifiers, and – where available – Wikidata identifiers, organized by volume.

2. **Bibliographic tables.** The same seed records used by the KG population pipeline (Section 7.5), indexed so that the backend can reconstruct the full FaBiO bibliographic chain for any given paragraph at query time.

3. **Pre-computed explorer subgraphs.** Each encoding a neighborhood of the KG as a node–edge list, together with a top-level manifest enumerating the available subgraphs and their metadata.

4. **Facsimile page images.** Scans of the 1912 Medici Society edition, organized by volume, enabling side-by-side comparison of the digital text with its print source.

Users enter the application through a landing page that frames the project rather than launching directly into a single view, namely a hero statement summarizing the dissertation, four headline statistic groups (GT coverage, ObliquER mention and entity counts, KG triple count, and annotation-corpus mention count), a featured-biographies gallery of the 16 GT biographies organized by the six de Vere volumes from which they are sampled (rendered as cards with

[33] FastAPI: `https://fastapi.tiangolo.com/`
[34] OpenAPI: `https://learn.openapis.org/specification/`.

facsimile-thumbnail previews), and short teasers for the main components reachable from the top navigation bar – the annotation viewer, the ontology documentation, the KG explorer, the knowledge-base browser, the ObliquER demonstrator, the SPARQL endpoint, and the publications and about pages. From the gallery, two reading paths diverge. The biography paragraph viewer (Section 7.7.2) is the central reading interface, presenting a single paragraph in a split-pane layout against its facsimile source. The knowledge-base browser (Section 7.7.3) is the corresponding entity-first path, exposing a paginated listing organized by class; each item resolves to a fragment-routed entity page whose URL fragment is the entity's IRI in the populated graph, so any link copied from the site is directly resolvable as a Linked Data subject.

The KG explorer (Section 7.7.4) supports exploratory navigation rather than sequential reading, rendering interactive force-directed views over pre-computed subgraphs of the full populated graph.

For users who require direct access to the underlying data, a browser-based SPARQL editor (Section 7.7.5) forwards queries to the triple-store endpoint, and the ontology documentation (addressing RQII.32 for ontology documentation access, cf. Table 8.13), generated from the RDF/XML source by WIDOCO,[35] is mounted as a static portal. A small set of project-level views – a publications list, an about page, the ObliquER landing page, and the pipeline demonstrator (Section 7.7.6) – round out the site. Assets and facsimile page scans are served as static files.

7.7.2 Biography paragraph viewer

The biography viewer is the application's central interface (addressing RQII.30 for inline provenance and RQII.31 for links to source text passages, cf. Table 8.13) and is designed around a split-pane layout that juxtaposes the source document with its computational analysis. Its left pane displays the facsimile page image corresponding to the current paragraph; its right pane presents the digitized paragraph text with inline annotation highlights.

Mention rendering and overlap resolution. Each mention produced by the GT annotation is rendered as a color-coded inline span within the paragraph text. Four CSS classes correspond to the four mention types defined in the Viewsari ontology:

[35] WIDOCO ontology documentation toolkit: `https://github.com/dgarijo/Widoco`.

Mention type	CSS class	Color
Explicit artwork mention	`explicit`	amber
Implicit artwork mention	`implicit`	blue
Coreferent	`coref`	purple
Generic mention	`generic`	gray

Because there may be overlapping mention spans, particularly when a coreference chain and another mention share a substring, the backend applies a priority-based overlap resolution algorithm before rendering. Mentions are sorted by start offset and, for ties, by span length (longer first).

Mention detail panel and Wikidata integration. Clicking an annotated span opens a detail panel that displays the mention's metadata, namely the mention type, surface form, entity identifier, and, if the entity was successfully linked, a clickable Wikidata URI. For linked entities, the application issues a client-side request to the Wikidata REST API to retrieve the entity's thumbnail image (property `P18`), inception date (`P571`), collection (`P195`), and material (`P186`). These are displayed in a structured card alongside the mention, providing immediate contextualization without requiring the user to leave the Viewsari interface.

Entities flagged as OOKB are visually distinguished by a dashed-border variant of their mention-type color, signaling to the user that the entity lacks a Wikidata counterpart.

Bibliographic chain display Below the annotation panel, a breadcrumb-style bibliographic chain reconstructs the FaBiO hierarchy for the mention and paragraph.

Facsimile navigation The facsimile pane supports page-level navigation across the entire biography. A page map, computed at startup by scanning all paragraphs for their printed page numbers, enables bidirectional linking between paragraph identifiers and facsimile images. When the user scrolls to a new paragraph, the facsimile pane automatically advances to the corresponding page; conversely, clicking a page thumbnail in the navigation bar scrolls the text pane to the first paragraph on that page.

Vasari speech feature As an additional interpretive aid, the viewer includes a "Vasari Speech" sidebar that generates playful first-person commentary in Giorgio Vasari's voice for each explicit and implicit mention. These speech bubbles are selected deterministically from a curated pool of ten templates per mention type, with the surface form interpolated into each

template. The templates were generated with Claude, given the input texts from The Lives. The feature is intended to make the distinction between explicit and implicit mentions immediately tangible. For explicit mentions, Vasari proudly names the work, while implicit mentions produce speech bubbles that reflect on the art of allusion.

7.7.3 Knowledge-base browser

The knowledge-base browser (Browser & explore) (addressing RQI.8 for index search across extracted resources, cf. Table 8.13) exposes the entity layer of the populated graph as a paginated, server-rendered listing, with one tab per top-level entity class. Entities are rendered as cards that display the canonical label, alternative labels harvested from Wikidata, the Wikidata link where available, and a free-text search filter. Artworks are paginated by an asynchronous endpoint that supports filtering by source category (GT, ObliquER, or all) and free-text query; each card shows a source-category tag, the volume and paragraph in which the entity is first introduced, and a badge indicating whether the entity is linked to Wikidata or marked OOKB. The ObliquER-mentioned listings are paginated identically and accept an optional strategy parameter that allows the user to restrict the listing to one of the prompt variants in Section 7.4.8.4.

The entity-detail pages use fragment-based routing, with the server returning a single template that loads the entity record via a client-side fetch and renders the entity's full property table together with a list of its textual occurrences, each carrying a deep link back to the source paragraph in the biography viewer (Section 7.7.2). This routing pattern ensures that the entity interface displayed in the address bar is identical to the entity's IRI in the populated graph (under the Viewsari knowledge-base namespace), so that copying a URL from the website yields a citable, persistent identifier directly resolvable. The browser is the most direct expression of the principle that the website is a view onto the KG rather than an alternative to it, since every browser page corresponds, one-to-one, to an instance in the Turtle serialization, and the same entity can be reached either through the website's interface or through a SPARQL query against the triple-store endpoint.

7.7.4 Knowledge graph explorer

The KG explorer (the dominant locus for the visualization requirements RQI.9, RQI.10, RQI.11, and RQI.12 of Table 8.13 – semantic and visual search, open-ended browsing, visibility of connection strength, and visualization of resources) provides an interactive force-directed graph

visualization of pre-computed subgraphs from Section 7.5. It is implemented with D3.js v7 (Bostock et al. 2011).[36] The subgraph-fetch endpoint returns a specific subgraph by seed entity and mode (person, artwork, or biography), with matching proceeding from exact label match to partial match to cross-mode node search; and an autocomplete-suggestion endpoint that provides real-time suggestions as the user types, drawing from the deduplicated label set of all loaded subgraphs. In the typology of Section 5.2.6.2, this view occupies the same role as the InTaVia entity landing page (Section 5.2.6.2), which renders an entity's ego-network as the primary mode of access to the surrounding KG (Windhager et al. 2024); the chronological-timeline component shown in Figure 5.13 is reserved as future work pending the integration of dated historical events into the populated graph.

Graph rendering. Nodes are rendered as rounded rectangles for persons and biographies and as small circles for co-occurrence nodes. A color palette distinguishes the four node types – person, artwork, co-occurrence, and biography – and each node displays a truncated label together with a type-specific icon badge. Edges are drawn as straight lines with arrowhead markers and optional dashed styling for secondary relationships such as part-of relations or sibling artwork links, and edge labels are positioned at the midpoint of each link.

The force simulation combines a *link force* that keeps co-occurrence triads compact while leaving room for other relationships; a *charge force* with type-dependent strength that prevents node overlap; a *center force* that draws the graph toward the viewport midpoint; a *collision force* that provides a minimum separation between node bodies; and an *axis-aligned force* that gently separates nodes along the vertical axis by type, placing persons above artworks above paragraphs. After the simulation settles, an automatic zoom-to-fit transformation centers and scales the graph within the viewport.

Wikidata image enrichment. For every node that carries a Wikidata identifier – both artwork and person nodes – the explorer renders a circular image thumbnail adjacent to the node. The implementation proceeds in two phases. During the initial graph layout, each node with a Wikidata identifier receives a circular placeholder with a colored border matching the node type; after the graph has rendered, the application collects the distinct identifiers, batches them into requests to the Wikidata API, and asynchronously replaces each placeholder with a thumbnail derived from the Wikidata image property of the corresponding entity.

The result is that artwork nodes display a thumbnail of the artwork itself, while person nodes display a portrait. Nodes whose Wikidata entity lacks an image claim retain the placeholder.

[36] D3.js and updated versions: `https://d3js.org/`.

Interaction model. The explorer supports the following user interactions:

- **Click** on a node to highlight its incident edges and display a detail panel listing the node's type, label, connection count, and a clickable Wikidata URL.
- **Double-click** on an artwork or paragraph node to navigate to the corresponding paragraph in the biography viewer.
- **Drag** any node to reposition it; the force simulation resumes temporarily to adjust neighboring nodes.
- **Zoom and pan** via mouse wheel and drag on the background.
- **Search** via the toolbar input with real-time autocomplete; selecting a suggestion loads the matching subgraph.
- **Preset buttons** for frequently explored entities (Giotto, Michelangelo, artists, artworks, biographies) provide one-click access to representative subgraphs.

For person nodes, an additional "Explore neighborhood" button triggers a search for the selected person's name and loads the corresponding person-centered subgraph.

7.7.5 SPARQL query interface

The SPARQL view (addressing RQII.33, the explicit SPARQL-interface requirement of Table 8.13) provides a browser-based query editor with a tabular result display. Queries submitted through the interface are forwarded to the same triple-store endpoint described in Section 7.6.1. A set of pre-defined example queries is offered, covering representative CQs from the requirements catalog (Section 7.3.1.1):

- Retrieval of all persons mentioned in a given biography.
- Listing of all co-occurrences involving a specific person.
- Enumeration of paragraphs containing artwork mentions linked to a given Wikidata entity.
- Traversal of the full bibliographic chain from a mention to the edition node.
- Provenance queries tracing an annotation back to its generating activity.

The SPARQL interface primarily serves as an expert tool for directly exploring the ontology and populated KG. Also, it allows for an ad hoc verification of the CQs against the populated graph.

The provision of a curated example library, rather than an empty editor, follows the convention established by the Wikidata Query Service and related SPARQL-driven exploration tools surveyed in Section 5.2.6.1, which similarly offer a corpus of seed queries that users can adapt rather than author from scratch.

7.7.6 ObliquER pipeline demonstrator

The application includes an interactive demonstration view that illustrates the ObliquER extraction pipeline (Section 7.4) without requiring a live connection to an LLM. The view is designed to make the extraction process legible to users who may not be familiar with prompt engineering or generative IE.

A prominent disclaimer banner clarifies that the page is a demonstration only and no live LLM is invoked; the displayed annotations are precomputed GT results from the actual extraction run. The prompt text and commentary are provided for illustrative purposes.

Interface layout. The view is organized as a split-pane layout. The left pane presents three elements in sequence:

1. **Passage selector.** The user may choose between two representative passages from the GT corpus (Section 7.4.4): a passage from the Life of Cimabue describing the Maestà panel for Santa Trinita (11 mentions), and a passage from the Life of Giotto describing his burial in Santa Maria del Fiore (3 mentions). The two passages were selected to illustrate different densities and distributions of mention types.

2. **System prompt.** A syntax-highlighted rendering of the actual ObliquER few-shot system prompt, showing the role description, task formulation, mention-type taxonomy (*explicit*, *implicit*, *coreferent*), and output schema. The prompt text is reproduced verbatim from the static content used in the extraction run.

3. **Input passage.** The paragraph text as it would be submitted to the LLM. When the user triggers the extraction, the text is re-rendered with inline annotation spans that are progressively revealed.

The right pane displays the extraction output:

1. **LLM commentary.** A narration bar styled as a first-person commentary in a deliberately self-important register (e.g. "*Ah, Cimabue and his monks of Vallombrosa! Let me examine this passage carefully. . .* "). The commentary updates at intervals as mentions are revealed,

drawing attention to noteworthy features of the extraction – such as the density of implicit references or the structure of coreference chains. The tone is intentionally informal and serves a pedagogical rather than analytical function, making the distinction between mention types immediately tangible to a non-specialist audience.

2. **Mention cards.** For each detected mention, an animated card appears showing the mention type (color-coded as in the biography viewer), the surface form, the entity identifier, the coreference antecedent (for coreferent mentions), the resolved label, and – where available – a clickable Wikidata link. Entities flagged as OOKB carry a red badge.

Animated extraction sequence. Pressing the "Run ObliquER extraction" button initiates a scripted animation sequence that simulates the temporal progression of a generative extraction pass:

1. The passage text is re-rendered with hidden annotation spans.
2. After a brief delay, the LLM commentary bar appears with an introductory remark.
3. Each mention is then revealed in document order: the corresponding inline span transitions from transparent to its type-specific highlight color (amber for explicit, blue for implicit, purple for coreferent), and a result card slides into the right pane.
4. The commentary text updates at intervals, providing running observations.

The purpose of this animation is to convey the sequential, left-to-right nature of the extraction process, in which each mention is identified in context and linked to previously identified entities. Furthermore, it provides a visual demonstration of the mention-type taxonomy, since the progressive color-coding makes it immediately apparent which references Vasari names explicitly and which he leaves implicit.

Rationale behind the demonstrator. The ObliquER demonstrator is intended as a communication tool for interdisciplinary audiences – particularly art historians and DH researchers who interact with the extraction outputs. By presenting the prompt, the input, and the structured output side by side, the view makes the "black box" of generative extraction transparent, allowing the user to follow what ObliquER does, what instructions the model receives, what text it processes, and what structured annotations it produces.

7.7.7 Design principles

The web interface is structured around three foundational principles, each articulating a specific methodological commitment. These principles directly address a persistent interdisciplinary asymmetry, in that a significant epistemic gap separates the technical infrastructure of Viewsari, including SPARQL querying, ontology engineering, and generative entity extraction, from the established practices of art-historical domain experts who form the primary user base. Throughout the project, domain experts have articulated substantive concerns about ObliquER and the broader entity-extraction pipeline, particularly the risk that interpretively constructed outputs from Vasari's text could be presented to users as established fact. Consequently, the interface must not only expose data but also make explicit the conditions under which it was produced, distinguishing between structural givens, inferences generated through extraction, and the limits of confident assertion. The principles are designed to operationalize this requirement, rendering methodological commitments accessible to domain experts who may not possess technical fluency.

Ontological transparency. Each data-bearing element within the interface is explicitly mapped to a corresponding class, property, or individual in the Viewsari ontology. The bibliographic breadcrumb reflects the `frbr:isPartOf` relation; mention color coding encodes the `viewsari:mention` type hierarchy; and explorer node types correspond to the ontology's top-level entity classes. This systematic alignment renders the KG's ontological structure transparent, enabling users to apprehend the underlying commitments without direct engagement with RDF syntax.

All about the provenance. The interface operationalizes the distinction between structurally grounded triples (bibliographic layer) and interpretively grounded triples (extraction layer) by separating the facsimile view, which presents the document as an object, from the annotation overlay, which presents the document as an interpreted source. The visual marker for OOKB entities delineates the boundary between external KBs and Viewsari-internal entities, thereby manifesting the provenance-aware design articulated in Section 2.4.

Layered learning. The adoption of a layered architecture in the Viewsari user interface corresponds to the view types of the Sampo model – landing, perspective, and entity pages (see Section 5.2.6.2), reconceptualizing the hierarchy as a continuum of technical expertise rather than a predetermined sequence of views (Hyvönen 2023, Ikkala et al. 2022). Each component aligns to a distinct level of analytical granularity. The homepage presents a corpus-level overview; the

biography viewer supports paragraph-level analysis; the mention panel exposes entity-specific metadata; and the SPARQL interface affords unrestricted graph-level querying. By structuring the interface this way, users can engage with the KG according to their technical proficiency and research objectives. The design obviates the need for prior familiarity with the underlying ontology – its schema is, ironically enough, implicitly interwoven into the interface.

7.7.8 Continuous updates

We maintain the companion web application as a living artifact. This means that, given the work-in-progress status of, for example, the KE steps with ObliquER, we continuously revise and repopulate the KG, which may lead to differences in numbers between the written work and the deployed UI. This also affects the SPARQL query interface.

Worked example: federated enrichment via Wikidata. A representative instance of this continuous development is the extension of the SPARQL interface (Section 7.7.7) with *working* federated queries against the Wikidata Query Service. This directly operationalizes `CQII.27` ("Can I perform a federated SPARQL query combining Viewsari data with an external endpoint?"), which the competency-question catalog had previously recorded only as an illustrative, non-executable template. The deployed interface now ships two ready-to-run examples that join Viewsari entities to Wikidata live via a `SERVICE` clause and surface, for a Wikidata-linked artwork or person, externally held attributes such as creator, holding collection, inception date, or biographical dates. In doing so, it also satisfies the corresponding requirements `RQII.22` (federated SPARQL across linked datasets), `RQII.20` (enrichment of entities from *The Lives* with information from Wikidata and similar authorities), and `RQI.7` (federated queries for fact-checking across external datasets), and it provides a concrete answer to the linked-data-enrichment cluster of CQs that ask which entities carry external authority links and which properties are mapped (`CQI.13`, `CQII.13`, `CQII.25`), as well as the enrichment-capability question `CQII.18`.

8 Results

This chapter systematically documents the outputs generated in alignment with the conceptual and empirical contributions outlined in Chapter 1.[1] The evaluation of these outputs against formal criteria follows in Chapter 9.

In the following, we report the four components:

1. The **Viewsari ontology** in its final instantiated state,
2. The **ObliquER** extraction and linking results over the Vasari ground truth corpus and ArtPedia,
3. The **populated KG** statistics, and
4. The **Viewsari web interface** in its final deployed form.

8.1 The Viewsari ontology: final state and instantiation

The result reported in this section is empirical contribution **E1**, which we theorized in Chapter 1. The Viewsari ontology, an OWL 2 DL ontology, operationalizes the conceptual commitments **C1**–**C3** as an annotation- and provenance-model grounded in textual mention spans and computational extraction processes. Concretely, the ontology implements the three-layer architecture (bibliographic, structural, content) that **C1** proposes as the representational response to the pre-interpretive entity assumption, integrates PROV-O (Lebo et al. 2013) and the Web Annotation Ontology Sanderson et al. (2017) as the backbone for the hermeneutic-provenance commitment of **C2**–**C3**, and represents extracted entities together with the conditions of their production: extraction agents, prompt templates, and the mention typology (explicit, implicit, coreferent,

[1] The ontology and requirements-engineering results reported here extend the iterative design phases first published in Ondraszek et al. (2024a).

generic) that records the textual basis of every extraction. The output of **E2** (ObliquER) and **E3** (the populated KG and web interface) is reported in Section 8.2, Section 8.3, and Section 8.4.

Combined with the output from ObliquER, the Viewsari ontology forms the Viewsari KG, the core output of this dissertation. All development steps were evaluated iteratively and in collaboration with domain experts. This includes concept and prototype-based validation grounded in the KG. Competency questions test the ontology's accuracy in supporting queries relevant to domain experts; this evaluation is reported in Section 9.1. A reasoner test that combines the TBox with a minimal ABox subset covering all classes and predicates used in the populated graph is available in the project repository and is suitable for offline classification with HermiT (Section 7.6.4).

The iterative design process that produced this architecture is documented in Section 7.3.1 and Section 7.3.2; the figures below show its final form. Figure 8.1 shows the bibliographic layer, with four fixed individuals, anchoring FaBiO Peroni and Shotton (2018) from the abstract work down to dereferenceable web representations of every volume, biography, and page. Figure 8.2 shows the content layer, organized around `viewsari:extracted_content` as a uniform PROV-typed superclass. The persons, artworks, organizations, co-occurrences, and the mention taxonomy (explicit, implicit, coreferent, generic, with two artwork-specific specializations) all share this provenance chaining, and typed activity classes record the pipeline stage that generated each instance. Figure 8.3 shows the two layers joined through the structural layer (`doco:Paragraph`, so that every extracted entity is traceable back to a specific paragraph in a specific edition and a specific web manifestation.

Table 8.1 lists the resulting classes, object properties, and datatype properties organized by layer. Elements introduced in the current iteration relative to the first formalization are marked with †.

8.2 ObliquER: entity recognition and linking results

The ObliquER pipeline and its outputs constitute contribution **E2** of Chapter 1. As a modular generative extraction pipeline for implicit and explicit entity recognition and linking, it implements **C2** in practice by recording each entity extraction step as a PROV-O activity in the resulting graph.

8.2.1 Final prompt templates

Table 8.2 inventories the five prompt templates used in the ObliquER pipeline, indicating for each whether its output is ingested into the deployed Viewsari KG (Section 7.5.5). Their design rationale, content blocks, and few-shot examples are described in Section 7.4.8.3, Section 7.4.8.4

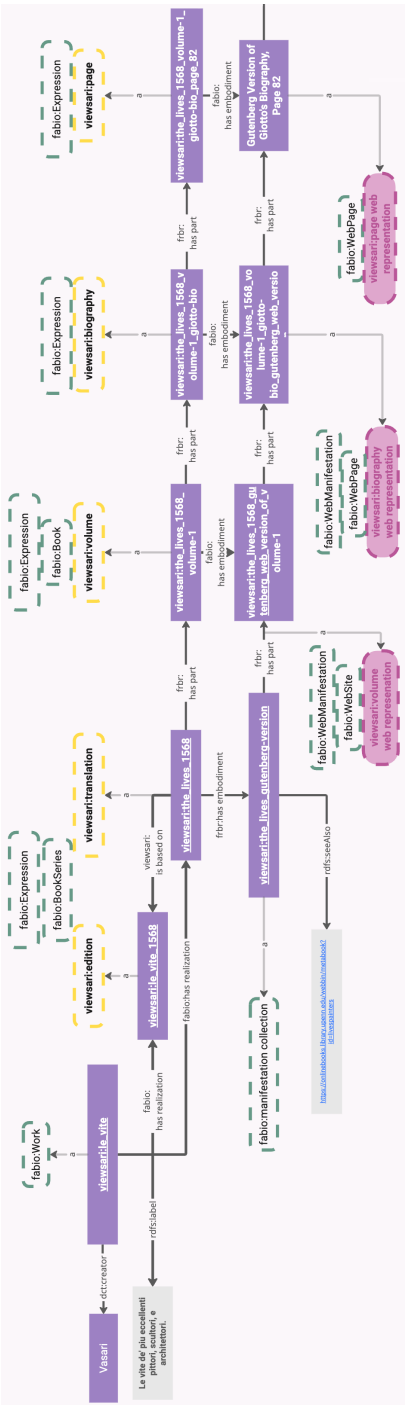


Figure 8.1: Final form of the Viewsari bibliographic layer. The FaBiO/FRBR work *Le Vite* is connected to its 1568 Giunti edition, the 1912 de Vere English translation, and the Project Gutenberg manifestation collection, with dereferenceable web representations at the volume, biography, and page levels.

Table 8.1: Summary of Viewsari ontology elements by layer in the final version. New elements introduced in the current iteration relative to the first formalization are marked with †.

Layer	Element	Purpose
Bibliographic	`fabio:Book / viewsari:le_vite`	Abstract work individual
	`viewsari:edition`	Published version (e.g., 1568 Giunti)
	`viewsari:translation`	de Vere 1912 English translation
	`viewsari:volume`	Physical/digital volume unit
	`viewsari:biography`	Artist biography, as a discourse unit also part of the structural layer
	`viewsari:volume web representation`†	Web manifestation of a volume
	`viewsari:biography web presentation`†	Web page hosting a biography
Structural	`viewsari:page`	Page-level structural node
	`viewsari:page web representation`†	Individual page anchor
Content	`viewsari:extracted content`†	Container for entities and mentions extracted from text
	`viewsari:person`	Artists, patrons, figures
	`viewsari:artwork`†	Artwork entities (explicit and implicit)
	`viewsari:organization`†	Guilds, institutions
	`viewsari:cooccurrence`	Person–person co-occurrence
	`viewsari:mention`†	Shared superclass for all mention types
	`viewsari:explicit artwork mention`†	Canonical title present in text
	`viewsari:implicit artwork mention`†	Recoverable artwork, no canonical title
	`viewsari:coreferent`†	Back-reference to earlier mention
	`viewsari:generic mention`†	Non-entity discourse reference
Provenance	`viewsari:named entity recognition`†	Named entity recognition activity subclass
	`viewsari:entity linking`†	Entity linking activity subclass
	`viewsari:cooccurrence analysis`†	Co-occurrence extraction activity

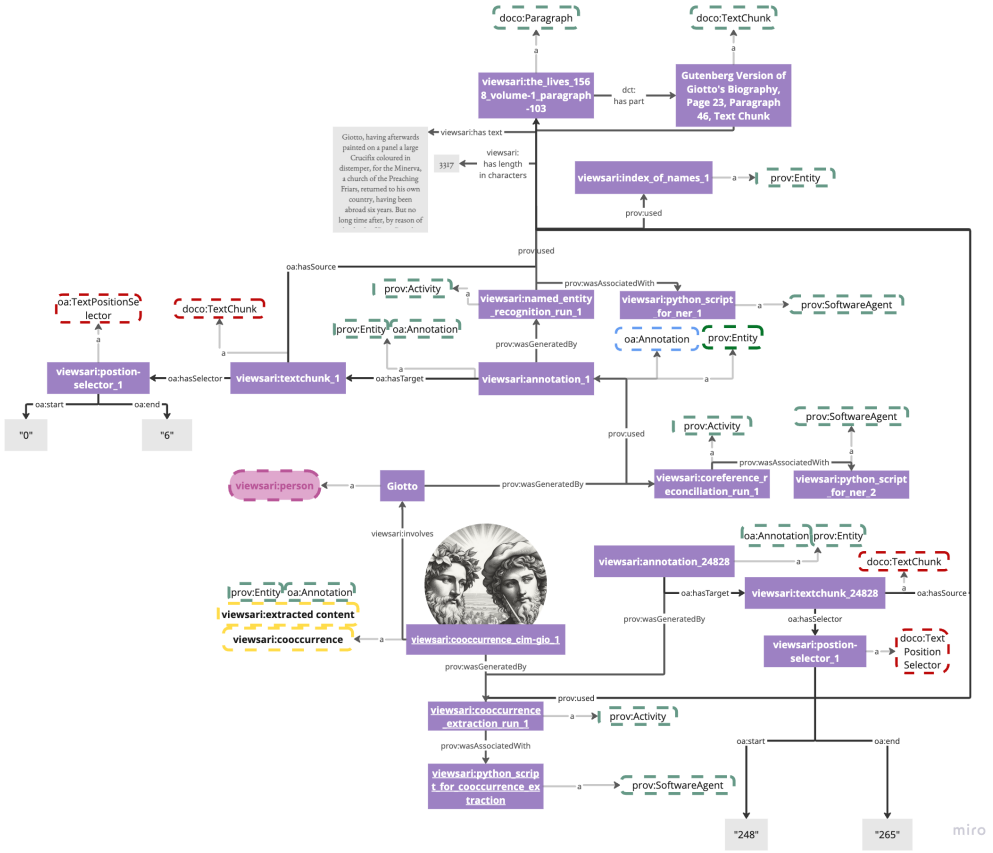


Figure 8.2: Viewsari content layer, with `viewsari:extracted_content` as the uniform PROV-typed superclass over persons, artworks, organizations, co-occurrences, and the mention taxonomy. Three typed activity classes (`named_entity_recognition`, `cooccurrence_analysis`, `entity_linking`) record the pipeline stage responsible for each instance.

(few-shot and ontology-guided recognition variants), Section 7.4.8.5 (chain-of-thought recognition variant), and Section 7.4.9.1 (entity linking). The further ablation prompts (A1–A4) that probe the task-decomposition, domain-content, ontology-grounding, and demonstration-count axes are described in Section 7.4.8.4, empirically compared in Section 9.4.6, and cataloged with their source files in Section A.0.0.0.2, where verbatim template listings of all variants are also reproduced.

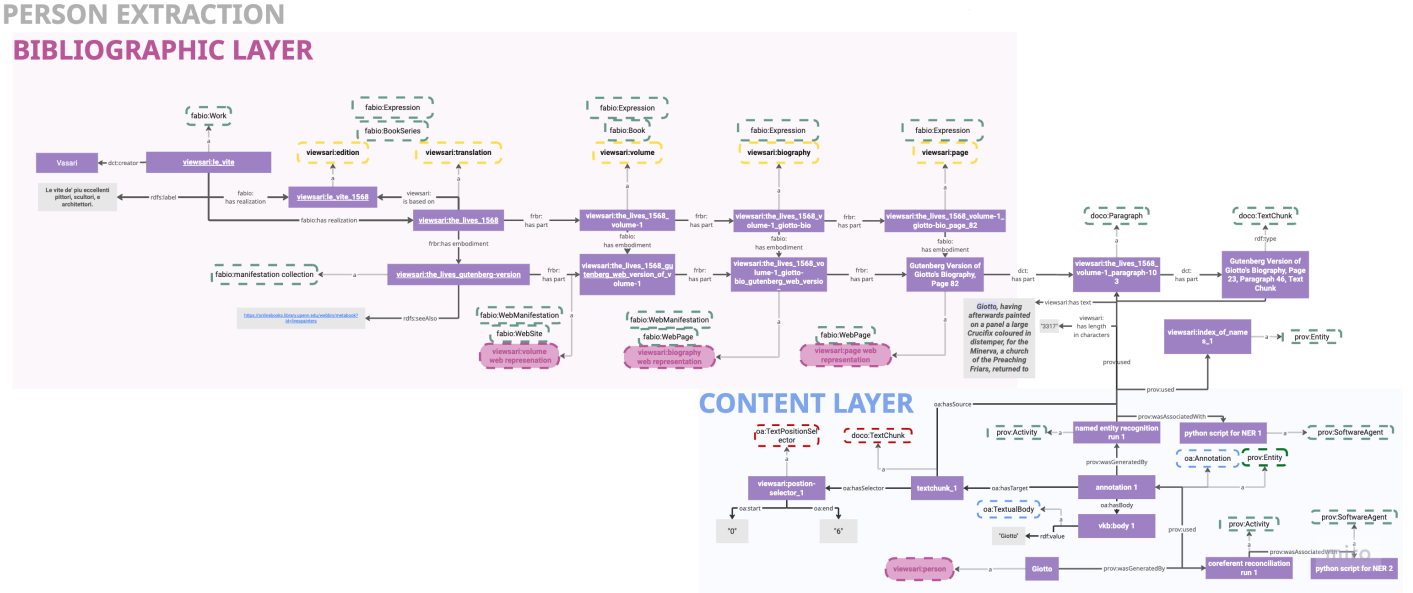


Figure 8.3: Combined view of the content and bibliographic layers, joined via the structural layer. Every content-layer entity is anchored to a `doco:Paragraph`, which in turn carries page-extent properties pointing into the bibliographic layer's web representations, so that every extraction is traceable to a specific edition, translation, and online source page.

Table 8.2: ObliquER prompt template inventory and KG deployment status. *In KG* indicates whether the variant's output is loaded into the deployed Viewsari KG (Section 7.5.5); variants marked *No* are retained in the project repository for the empirical comparison in Chapter 9 but do not contribute mentions or links to the populated graph.

Template	Task	Ontology	In KG	Key feature
ER few-shot	Recognition	No	Yes (parallel)	Baseline; single extended example
ER chain-of-thought	Recognition	No	No	4-step reasoning scaffold; provenance trace
ER few-shot - ontology-guided	Recognition	Yes	Yes (canonical)	KG modeling example; ontology-aligned vocabulary; $k = 2$ demonstrations
EL - few-shot	Linking	No	Yes	ranker over retrieved candidates; 6 few-shot examples (5 Michelangelo + 1 NIL)

8.2.2 Ground truth and ObliquER extraction statistics for the Viewsari dataset

This subsection summarizes the results from manual annotation and the ObliquER pipeline on the Viewsari corpus, serving as the descriptive counterpart to the evaluation reported in Chapter 9. The full per-biography breakdown of the ground truth has been provided separately in Section 7.4.7.

ObliquER ground truth statistics. The Viewsari ground truth comprises $N = 232$ annotated paragraphs across 16 biographies, sampled from six volumes of the de Vere translation.[2] The negative fraction is computed over this corpus; the heuristic flag for negative examples marks paragraphs likely to contain no artwork mention.

The GT corpus is described in detail in Section 7.4.7; Table 7.11 there gives the per-biography mention-type breakdown. The corpus-level totals – 16 biographies, 232 paragraphs, 2,422 mention annotations across the four Viewsari mention types, and 848 globally-deduplicated artwork entities – are detailed there; Table 8.4 below instead sets the GT layer *as deployed* (2,439 mentions) beside the ObliquER extraction layer and their union in the populated graph.

[2] The original stratified-sampling design (Figure 7.25) targeted 221 paragraphs with a soft cap of 11 % per biography. The annotated corpus realized after quality control contains 232 paragraphs; the additional 11 paragraphs are admitted as a result of the dev-set carve-out described in Section 7.4.7, in which examples used in any few-shot prompt remain in the corpus but are excluded from the evaluation partition.

Table 8.3: Corpus-level statistics for the Viewsari ground truth ($N = 232$ paragraphs, 16 biographies).

Volume	Contents (period)	Paragraphs	Neg.
Vol. I	Proto-Renaissance (Trecento)	36	5
Vol. II	Early Renaissance (architecture)	16	1
Vol. III	Early Renaissance (painting/sculpture)	92	10
Vol. IV	High Renaissance	9	1
Vol. VII	Mannerism	15	1
Vol. IX	High Renaissance / Mannerism	64	4
Total		**232**	**22**

ObliquER extraction statistics over the full corpus. The ObliquER pipeline applied to all ten volumes of the de Vere translation produces a substantial mention layer. Across the 3,479 paragraphs, the canonical build (Run 2) contains **84,215** mentions typed under the four Viewsari mention subclasses.

The ObliquER layer of the canonical build aggregates across three prompt variants. Run 1 ingested two – few-shot and ontology-guided – for 55,246 mentions; Table 8.5 reports the per-variant raw output of the prompt-engineering experiment from which both were drawn. Run 2 then re-authored the ontology-guided prompt with new few-shot demonstrations and ingested its output as a third parallel layer (28,969 mentions), bringing the canonical ObliquER layer to the 84,215 mentions of Table 8.4, whose per-type composition the ObliquER column of that table records.[3] The fourth recognition configuration explored – chain-of-thought, with 23,602 raw mentions – is retained in the project repository but loaded into neither build (Section 7.5.5).

8.2.3 Ground truth and extraction statistics: ArtPedia

The ArtPedia dataset is used in this dissertation as a contemporary, cross-domain counterpart to the Viewsari ground truth, characterizing what is annotated rather than scoring what was extracted. The corresponding evaluation – transferability of ObliquER from sixteenth-century biographical prose to twentieth-century art-historical commentary – is reported in Chapter 9.

[3] The few-shot raw total of 28,058 in Table 8.5 is eleven mentions higher than the 28,047 actually loaded, a 0.04 % gap that persists unchanged in both builds. The eleven mentions belong to a single opening paragraph (volume 1, paragraph 1), whose few-shot response file was absent from the frozen data snapshot used for population and was therefore never loaded; the file is well-formed and parses without error, so a future rebuild can recover it.

Table 8.4: Composition of the deployed Viewsari KG by source. The *Ground truth* column reports the manually annotated layer *as deployed* in the graph: 2,439 `prov:Person`-attributed mentions. This is the *pre-correction* ground truth from which the graph was populated; the annotated corpus was *subsequently* corrected down to 2,422 mentions by merging 17 headless-coreferent chains into their antecedents (Section 7.4.7). Therefore, the GT layer deployed and reported in this table is different from the corrected corpus in Chapter 7 and will be reconciled in the next redeployment. The *ObliquER extraction* column is the recognition layer over all biography instances and paragraphs, and *KG (both)* is their union; all three columns report the canonical build (population Run 2). Mention rows are additive (ground truth + ObliquER = both); the biography and paragraph rows are not, because the ground truth is the 16-biography / 232-paragraph annotated subset of the full corpus. The artwork-entity rows are likewise additive: Run 2 links and clusters the recognized artworks of the re-authored ontology-guided variant (Section 8.3.5), so the ObliquER entity column is now populated rather than pending. The GT column reports the deployed pre-correction layer (852 globally-deduplicated artworks; the corrected corpus of Section 7.4.7 resolves to 848, and the per-biography-sum convention used in the accompanying conference paper yields 956). The remaining two recognition variants – few-shot and the original ontology-guided – contribute mentions but are not yet clustered, so the ObliquER entity counts are an upper bound that a global reconciliation pass will reduce (Section 8.3.2). Section 6.4.2 reports 168 biographies as the source-corpus scope; of these, 165 are materialized under `viewsari:biography`, the 3-unit gap reflecting supplementary entries in the de Vere translation (the academicians-of-design and festive-preparations appendices and Vasari's self-life in volume 10).

Statistic	**Ground truth** *(as deployed)*	**ObliquER extraction**	**KG (both)**
Biographies	16	165	165
Paragraphs	232	3,479	3,479
Mention annotations (total)	2,439	84,215	86,654
Explicit artwork mention	221	14,296	14,517
Implicit artwork mention	811	39,856	40,667
Coreferent	1,094	21,189	22,283
Generic mention	313	8,874	9,187
Artwork entities (globally deduplicated)			
Unique artwork entities	852	14,952	15,804
Wikidata-linked	311 (36.5 %)	2,964 (19.8 %)	3,275 (20.7 %)
OOKB	541 (63.5 %)	11,988 (80.2 %)	12,529 (79.3 %)

Table 8.5: Raw mention counts per prompt variant from the Run 1 ObliquER recognition stage over all 10 volumes of the de Vere translation (`openai/gpt-oss-120b`). Counts are aggregated from the per-paragraph response files. These are the three prompt-engineering variants from which Run 1's ingested layer was drawn: the few-shot and ontology-guided variants as parallel mention layers, while the chain-of-thought variant is retained in the repository but loaded into neither build. Run 2 additionally ingests a re-authored ontology-guided variant (28,969 mentions; per-type composition in the ObliquER column of Table 8.4) not shown here, taking the canonical ObliquER layer to 84,215.

Variant	Total	Explicit	Implicit	Coreferent	Generic
few-shot (2 examples)	28,058	4,334	13,490	6,965	3,269
ontology-guided (2 examples)	27,199	5,037	12,802	6,385	2,975
Run 1 (raw) total	55,257	9,371	26,292	13,350	6,244
chain-of-thought (not ingested)	23,602	5,219	12,117	5,806	460

For integration with ObliquER, the original ArtPedia paragraphs were reconstructed from their Wikipedia source texts and mapped to the same GT format used for Viewsari (Section 7.4.10). Of ArtPedia's 2,930 paintings, 2,679 (91.4 %) could be reliably matched to a Wikipedia source paragraph at the time of harvesting; the remaining 251 were excluded. Canonical titles in the matched paragraphs identify *explicit* mentions, and all remaining sentences are annotated as *implicit*; coreferent and generic mention types are not used in the ArtPedia split. This produces a synthetic implicit environment that is structurally analogous to Viewsari but operates at a much larger scale and with substantially higher Wikidata coverage of the underlying entities.

Table 8.6: Mention and entity statistics for the ArtPedia ground truth as used in this dissertation (Wikipedia-paragraph reconstruction of the dataset). Annotations are restricted to *explicit* and *implicit* mention types; coreferent and generic annotations are not used in this split. OOKB counts are reported as percentages of the total number of entities and mentions.

Statistic	ArtPedia ground truth
Paragraphs	7,618
Paintings retained from ArtPedia (of 2,930)	2,679 (91.4 %)
Mention annotations (total)	21,774
Explicit artwork mention	2,222
Implicit artwork mention	19,552
Unique artwork entities	2,679
OOKB (% of entities)	0.26 %
OOKB (% of mentions)	0.27 %

Compared with the Viewsari ground truth in Table 8.4, ArtPedia is more than an order of magnitude larger (7,618 paragraphs and 21,774 mentions) than Viewsari's 232 paragraphs and 1,010 mentions, making ArtPedia a substantially heavier transferability test for the recognition pipeline. Furthermore, given its origin from Wikipedia, the OOKB ratio is inverted. Where the Viewsari ground truth records 63.4 % of entities as OOKB (on the globally-deduplicated convention used in Table 8.4), ArtPedia records only 0.26 %. Its Wikidata coverage is, by definition, near-complete. The dataset, therefore, tests ObliquER's recognition process at scale while holding the long-tail-linking challenge in abeyance, so that the linking-side results on ArtPedia (Chapter 9) isolate disambiguation quality from OOKB detection. The per-type breakdown of Table 8.6 – 2,222 explicit mentions against 19,552 implicit mentions – also makes explicit how strongly the ArtPedia split favors implicit references. Roughly 9 in 10 annotated mentions are implicit (≈9:1, against ≈4:1 in the Viewsari ground truth), a markedly higher imbalance that is precisely the property that motivates ArtPedia's use as a stress test for the implicit entity recognition.

ObliquER extraction over ArtPedia. The ObliquER recognition stage was run on the 2,679 ArtPedia painting records in two prompt configurations, a few-shot variant and a zero-shot variant. Table 8.7 reports the per-variant raw mention counts. Coreferent annotations are absent on both sides because the ArtPedia split does not use that mention type (Section 7.4.10).

Table 8.7: Raw mention counts from the ObliquER recognition stage over ArtPedia (2,679 paintings; `openai/gpt-oss-120b`). Counts are aggregated per-painting; parsing failures are excluded (9 of 2,679 in the few-shot variant; 92 in the zero-shot variant).

Variant	Total	Explicit	Implicit	Generic
few-shot	22,376	5,742	15,004	1,630
zero-shot	16,825	5,620	9,683	1,522

The two variants recover comparable explicit-mention counts (5,742 vs. 5,620) but diverge sharply on implicit detection (15,004 vs. 9,683), so the few-shot example contributes a 55% uplift on the implicit-recognition pathway that ArtPedia is specifically designed to stress. The full evaluation against the ArtPedia ground truth is reported in Section 9.4.7.

8.3 Knowledge graph population results

The figures reported in this section constitute the populated-graph half of empirical contribution **E3** of Chapter 1: a KG with full extraction provenance that makes the abstract claim of **C2** –

that provenance must remain visible at the point of use – concretely inspectable. All figures are computed by a fixed set of twelve SPARQL metric queries (M01–M12), executed against the canonical Turtle serialization of the closed KG. The twelve queries are partitioned into two categories along a single axis: *descriptive* metrics, which inventory what was populated, are reported in this chapter, while *evaluative* metrics, which test the methodology's quantitative claims (full provenance coverage, the 12–18 triples-per-mention upper bound, external linking coverage, web-manifestation reachability), are reported in Section 9.5.4 of the evaluation chapter.

Deployed and closed graph. The raw output of the population pipeline is the 2,448,545-triple dump. This is the artifact served by the public Viewsari SPARQL endpoint and consumed by the web interface (Section 8.4); throughout this dissertation, "deployed graph" refers to this unclosed serialization.[4] The closure rules are deliberately *not* applied at the deployed endpoint, so they do not collapse intermediate hops in the provenance walk, and the endpoint exposes users to the full `oa:Annotation` → `prov:Activity` → `prov:SoftwareAgent` chain.

For evaluation, the deployed graph is closed by applying the three SPARQL UPDATE rules of Listing A.2 (Section 9.1.3). Rule R1 propagates a paragraph anchor down the artwork → annotation chain, Rule R2 derives a direct cooccurrence → paragraph edge by parsing the cooccurrence label, and Rule R3 chains off R2 to give person → paragraph edges. The rules are idempotent and add 7,431 derived triples (5,855 artwork→paragraph edges from R1, since the entity-layer artworks materialized in this build did not yet carry the in-paragraph edge, plus 632 cooccurrence and 944 person edges from R2 and R3), taking the total to 2,455,976. The resulting serialization is called the *closed graph*, and all twelve metrics in Table 8.8 and Table 9.20 are evaluated against it. Where a number is meaningful only post-closure (M07, M12), the closed graph is the only valid frame of reference; where the comparison between the unclosed and closed views matters (M05's verbosity gap), the deployed-graph value is reported alongside.

Population Run 2: the current build of a living artifact. Every KG figure in this chapter and every KG metric in Chapter 9 (M01–M12) is computed over a single, coherent build of the populated graph, *population Run 2*, the current state of the graph. The Viewsari KG is a living artifact (Jiang et al. 2025): it is regenerated as the recognition and linking stages improve and as the ground truth is corrected, and each regeneration is paired with a full re-run of the metric suite

[4] The triple counts reported here are taken over the serialization with the recognition- and linking-prompt texts stripped to literal-free provenance stubs, so that the figure measures structural content and stays comparable to the earlier build (Section 8.3.5). The serialization actually served at the endpoint additionally inlines the prompt texts as provenance literals and is correspondingly larger at 2,559,459 triples; the prompt literals add no instances and leave every other metric unchanged.

rather than a piecemeal update of individual figures, so re-evaluation is a recurring step of the project rather than a one-off, consistent with the refinement-and-evaluation methodology surveyed by Paulheim (2016). We therefore report a coherent set of metrics from a single reproducible, provenance-stamped build and read every quantitative claim against the build on which it was computed. The previous build, population Run 1, is retained not as a parallel source of metrics but as a documented prior state against which the graph's growth is described (Section 8.3.5): every headline figure in this chapter is therefore a Run 2 figure, with the Run 1 value carried alongside only where the change is itself the point. Run 2 ingests three ObliquER recognition-prompt variants – the few-shot and ontology-guided variants carried over from Run 1 together with a re-authored ontology-guided variant with new few-shot demonstrations (Table 8.5) – and, for the first time, clusters and links the recognized artworks of that new variant into the entity layer, the step that Run 1 had left pending. Both builds were populated from the ground truth as it stood at population time, 2,439 human-annotated mentions (Table 8.4); the few-shot and the original ontology-guided recognition mentions inherited from Run 1 remain unclustered, so the entity-layer counts below reflect the GT and the new ontology-guided variant only.

We therefore report two figures for the human-annotated layer. The GT corpus is given at its corrected size of 2,422 mentions, because that is the artifact's current, cleanest state and the basis for all recognition and linking evaluation (Chapter 9); the KG is given at its populated size of 2,439 mentions, stemming from the GT layer as it stood at population time, before we handled different corrections – a layer that Run 2 carries over from Run 1 unchanged, so that the planned reconciliation to the corrected 2,422-mention corpus remains outstanding rather than realized (Section 7.4.7).

8.3.1 Population metrics: catalog

The metrics referenced throughout this have been selected based on different criteria:

(i) Ontology- and KG-evaluation literature. Task-based evaluation in the XD tradition (Presutti et al. 2009, Blomqvist et al. 2016, Carriero et al. 2021) treats competency-question answerability as the primary fitness-for-purpose metric, but is consistently paired with quantitative *structural* indicators – triple counts, instance counts per class, schema-utilization ratios – as the counterpart that confirms the ontology was actually populated against its intended use case (Section 5.2.1). The broader KG metrics survey of Hassan et al. (2025) catalogs their use across link prediction, classification, and end-to-end reasoning benchmarks. The Viewsari descriptive layer adopts this convention directly in M01–M04, which are the structural indicators reported across different XD-tested CH KGs, such as ArCo (Carriero et al. 2021). Provenance-coverage instruments (M06, M07) extend the structural-metrics axis with a PROV-O-specific reachability test that has no

analog in the pre-large language model literature but is mandatory for the hermeneutic-provenance commitment **C3** introduced in Chapter 1. The verbosity proxy (M05) operationalizes the triple-cost concern that the comparative modeling studies of CIDOC-CRM extensions consistently flag (Section 9.2).

(ii) The competency question catalog. Each metric that surfaces a coverage figure is anchored to at least one CQ in the Phase I or Phase II catalog (Section A, Section A); the practice of pairing each CQ with a SPARQL-resolvable answerability check follows the ArCo unit-testing methodology of Carriero et al. (2021). M09 (external linking coverage) is the population-side test of CQI.13 ("*Is person X linked to an external authority record (e.g., Wikidata)?*") and CQII.25 ("*Which entities in the KG are linked to Wikidata, and what properties are mapped?*") – both of which are unanswerable on an empty `owl:sameAs` graph regardless of ontology correctness. M10 (OOKB artworks) is the population-side test of CQII.8 ("*Which artworks in the KG are classified as OOKB entities – i.e., have no match in Wikidata?*") and CQII.26, and grounds the long-tail discovery argument that motivated Phase II in the first place. M12 (web-manifestation reachability) is the population-side test of CQII.21 ("*What is the direct web URL to the paragraph in the Project Gutenberg edition from which entity E or co-occurrence C was extracted?*"), which the requirements engineering of Section 7.3.2.1 identified as a fact-checking pre-condition for the digital-humanities personas.

(iii) The requirements catalog. The requirements table assembled in the first round of expert sessions (Table 7.2) names provenance, typed person-to-person relations, co-occurrence representation, contextual grounding, and external-source cross-referencing as the load-bearing commitments. Each row of the requirements catalog has at least one metric paired with it. M02 and M03 cover the typed-mention and co-occurrence requirements; M04 and M08 cover the provenance-activity requirement; M06, M07, and M09 cover the large language model-extraction provenance and Wikidata cross-referencing requirements (RQI.7, RQII.20) that Section 7.3.2.1 introduced under the C1–C3 framing; M10 operationalizes the OOKB-preservation requirement RQII.23; M12 operationalizes the direct-URL requirement RQII.17 (foreshadowed in RQI.4).

Table 8.8 lists the seven descriptive metrics reported in this chapter. Each row is keyed by the identifier used throughout the prose, the SPARQL file that produces it, and the subsection in which it is reported. Evaluative metrics are cataloged separately in Table 9.20 of Section 9.5.4.

Table 8.8: Descriptive population metrics: what is in the populated Viewsari KG. Each metric is computed by a single SPARQL query; the dissertation cross-reference column points to the subsection where the result is reported.

ID	Title	What it surfaces	Reported in
M01	Total triples	Triple count after closure.	Section 8.3.2
M02	Instance counts by ontology class	Materialized individuals per Viewsari class.	Section 8.3.2
M03	Mention taxonomy breakdown	Counts per concrete mention subclass.	Section 8.3.2
M04	Activity inventory	`prov:Activity` instances grouped by Viewsari activity class.	Section 8.3.3
M08	Activity → SoftwareAgent inventory	Pairing of each activity with its associated agent.	Section 8.3.3
M10	OOKB artworks	Count of artworks without a Wikidata `owl:sameAs`.	Section 8.3.4
M11	Authority namespaces in use	Distinct external authorities reached via `owl:sameAs`.	Section 8.3.4

8.3.2 Instance counts by ontology class

Table 8.9 summarizes the number of populated instances per core Viewsari ontology class following a complete pipeline run over the de Vere English translation corpus.[5]

The zero counts for `viewsari:organization` and `viewsari:location` reflect a deliberate Phase III scoping decision rather than a pipeline failure. Both classes are present in the ontology to support the planned extension beyond artworks, and their population is deferred to future work as discussed in Chapter 11.

Artwork mention count. Of **86,654 typed mention annotations**, **2,439** stem from the manually annotated ground truth and **84,215** were added by the ObliquER recognition stage.[6] These **84,215** ObliquER-recognition mentions are exactly the annotations that satisfy the strict

[5] The `prov:Entity` total is reported over the prompt-stripped serialization (2,448,545 triples): 134,875 = 112,113 annotations + 15,804 artworks + 6,958 prompt-provenance entities from the earlier extraction runs. This serialization deliberately drops the bulky prompt-text literals of the largest (final) extraction run – which account for the overwhelming majority of the raw output's size – to keep the served graph compact, while retaining the comparatively light prompt entities of the earlier runs; the asymmetry is recorded here for transparency. Querying the deployed graph, which keeps those prompt entities, returns 178,621 instead. All other instance counts are identical on both serializations.

[6] The populated GT count of 2,439 is the *pre-correction* figure, since the KG was populated from an earlier version of the ground truth, and only afterwards was the corpus corrected *down* to 2,422 mention instances by merging 17 headless-coreferent chains into their antecedents (Section 7.4.7). Run 2 carries this older, larger GT layer over from Run 1 unchanged and is stale relative to the corrected corpus; a re-deployment will reconcile it. The discrepancy is confined to the human-annotated layer and leaves the ObliquER recognition layer and every metric computed over it unaffected.

Table 8.9: Instance counts by Viewsari ontology class after population Run 2 (first stage + ObliquER second stage, with the new ontology-guided variant clustered into the entity layer) on the closed KG (2,455,976 triples). The Run 1 column carries the prior build for comparison (Section 8.3.5); rows in the bibliographic and structural layer are identical across the two builds.

Ontology class	Run 1	Run 2
Bibliographic and structural layer		
`viewsari:edition`	1	1
`viewsari:translation`	1	1
`viewsari:volume`	10	10
`viewsari:volume_web_representation`	10	10
`viewsari:biography`	165	165
`viewsari:biography_web_representation`	165	165
`viewsari:page`	3,026	3,026
`viewsari:page_web_representation`	3,026	3,026
`doco:Paragraph`	3,479	3,479
`doco:TextChunk`	83,154	112,113
`oa:Annotation`	83,154	112,113
Content layer (extracted entities)		
`viewsari:person`	443	443
`viewsari:artwork`	852	15,804
`viewsari:organization`	0	0
`viewsari:location`	0	0
`viewsari:cooccurrence`	541	632
`viewsari:mention` (any)	57,685	86,654
Content layer (provenance)		
`prov:Activity`	6,965	13,794
`prov:SoftwareAgent`	1	1
`prov:Entity` (inferred)	90,964	134,875

provenance-coverage metric (M07, Section 9.5.4). Each simultaneously carries a direct in-paragraph edge, a generating activity, and a typed `prov:SoftwareAgent`. The figure grows from Run 1's 55,246 by exactly the 28,969 mentions of the re-authored ontology-guided recognition variant added in Run 2.

Artwork entity count. Run 2 links and clusters not only the GT mentions but, for the first time, the recognized artworks of the re-authored ontology-guided variant. Together these resolve to **15,804 artwork entities**, of which **3,275 (20.7 %)** are linked to Wikidata via `owl:sameAs` and **12,529 (79.3 %)** are OOKB entities assigned stable Viewsari-internal IRIs (Section 8.3.4). The GT layer contributes 852 of these (311 linked / 541 OOKB); the new ontology-guided variant the remaining 14,952 (2,964 / 11,988); the few-shot and the original ontology-guided recognition layers carried over from Run 1 are retained in the graph as `oa:Annotation` nodes with surface-form bodies, full mention-type typing, and complete provenance, and only the re-authored ontology-guided variant is promoted to the entity layer through linking and clustering. Retaining the two superseded recognition runs rather than discarding them is a deliberate consequence of the central commitment of this dissertation that every extraction is an accountable, provenance-bearing artifact (Section 11.1). Deleting an earlier interpretive run to leave only the current best layer would reproduce the silent overwriting that the thesis argues against, so the older variants persist as full provenance annotation nodes that a query can still reach. The web interface makes this directly inspectable, since its artwork explorer lets a reader compare the recognition variants for the same work side by side. Clustering only the latest variant reflects that the re-authored ontology-guided run is the current recognition layer of record, not that the earlier runs are removed; their promotion to the entity layer is left to future work (Chapter 11). Two qualifications keep the 15,804 figure an *upper* bound. Cross-variant deduplication is incomplete: the 3,275 linked artworks collapse onto only 3,070 distinct Wikidata QIDs, and only one of the three recognition variants has been clustered as a consequence of the linking process, so a global reconciliation pass will reduce these counts rather than raise them. And the GT subset still rests on the pre-correction corpus: its 852 entities derive from the 2,439-mention layer, whereas the corrected 2,422-mention corpus resolves to 848 (one Wikidata-linked and three OOKB fewer), a reconciliation deferred to the next re-deployment.

The mention subclass distribution (Table 8.10) confirms the dissertation's claim that implicit references dominate Vasari's prose: 40,667 implicit artwork mentions stand against 14,517 explicit ones, a 2.80× ratio that the discussion of Section 8.2.2 attributes to the lack of naming standards and an orientation towards descriptive paraphrase over canonical naming. Coreferential mentions (pronouns, demonstratives) account for a further 22,283 annotations, and 9,187 generic mentions document phrases that name no specific recoverable entity.

Table 8.10: Mention taxonomy breakdown and instantiation count per mention type.

Mention class	Count
`viewsari:implicit mention`	40,667
of which `implicit artwork mention`	40,667
`viewsari:coreferent`	22,283
`viewsari:explicit mention`	14,517
of which `explicit artwork mention`	14,517
`viewsari:generic mention`	9,187
Total	86,654

The triples-per-mention ratio that this total implies (M05), and the interpretation of the closed total against the 12–18 upper bound used in the CIDOC-comparison section are evaluative and therefore deferred to Section 9.5.1.

8.3.3 Activity and agent inventory

Metric M04 counts `prov:Activity` instances grouped by Viewsari activity class; metric M08 pairs each activity with its associated `prov:SoftwareAgent`. The two together describe the run-time provenance footprint of the pipeline, namely which kinds of activities the population step produced, and which model executed them. The interpretation of these counts as evidence for the dissertation's full provenance claim is deferred to Section 9.5.4 – this subsection reports the inventory only.

Table 8.11: Activity inventory (M04) for population Run 2, with the Run 1 build for comparison. Each `prov:Activity` is typed by its Viewsari activity subclass. Whereas Run 1 recorded the linking pass as a single long-running corpus-level activity, Run 2's ontology-guided linking variant records entity linking *per paragraph*, raising the `entity_linking` count from one to 3,100; the two typed classes sum to 13,160, 634 short of the 13,794 `prov:Activity` total of Table 8.9, the difference being the 632 per-co-occurrence extraction activities together with the legacy foundation activities that carry no recognition/linking subclass.

Activity class	Run 1	Run 2
`viewsari:named entity recognition`	6,961	10,060
`viewsari:entity linking`	1	3,100
Total (typed)	6,962	13,160

8.3.4 External entity inventory and out-of-knowledge-base count

M10 counts artworks that lack a Wikidata `owl:sameAs` link (the OOKB set retained under stable Viewsari-internal IRIs), and M11 enumerates the distinct external authorities reached through `owl:sameAs` across the graph. Whether the resulting linked-vs-total proportions count as adequate coverage is an evaluative question; that question, together with the long-tail discovery argument, is taken up in Section 9.4.4.

Out-of-knowledge-base artworks (M10). Of the 15,804 populated `viewsari:artwork` individuals, 12,529 carry no Wikidata `owl:sameAs` link and are therefore retained in the graph under their Viewsari-internal IRIs. They remain first-class queryable nodes and continue to participate in mention, paragraph, and provenance relations; what they lack is an external reconciliation target.

Linking rate by mention type. The OOKB share is not uniform across mention types. Tracing each resolved artwork mention of the clustered ontology-guided variant (the only recognition variant promoted to the entity layer, Section 8.3.5) back to its entity's linking outcome, explicit, named artwork mentions resolve to a Wikidata identifier at **49.0 %** (2,416 of 4,928), against **30.2 %** (4,091 of 13,567) for implicit mentions; the same asymmetry holds in the human GT layer, where explicit mentions link at 84.9 % (185 of 218) against 31.9 % (245 of 768) for implicit ones. Implicit mentions therefore make up the bulk of the OOKB artwork population (9,476 of the variant's 11,988 OOKB artworks derive from implicit mentions). This per-mention-type gap is the mention-level counterpart of the entity-level OOKB share above and the mechanistic core of the canonization signature taken up in Section 11.1: the works Vasari names explicitly are disproportionately those that later acquired canonical identifiers, while the descriptively referenced works are disproportionately those that did not.

Authority namespaces (M11). We also report the number of distinct external authorities that can be reached from any individual in the graph via `owl:sameAs`. The graph is currently linked exclusively to Wikidata with 3,595 links (3,164 under the `https` and 431 under the `http` Wikidata namespace, jointly the 3,275 linked artworks and 320 linked persons of Table 9.21); reconciliation against other knowledge bases and authority files is registered in the ontology but not yet populated, in line with the Phase III scoping decision noted under Chapter 11.

8.3.5 Population history: Run 1 to Run 2

The canonical graph reported above is the second of two builds. Run 1 was the first deployed build, and it ingested the two recognition-prompt variants of the canonical recognition pass, few-shot and ontology-guided, and linked and clustered only the GT mentions into the entity layer. Run 2, the current canonical build, re-authors the ontology-guided prompt with new few-shot demonstrations, expanding the ingested set to three recognition variants, and, for the first time, materializes that variant's recognized artworks into the entity layer through the linking-and-clustering stage that Run 1 left pending (Chapter 11). All twelve metrics (M01–M12) and the evaluation of Chapter 9 are computed over Run 2; Run 1 is retained here only as a documented prior state, so that the direction and magnitude of the graph's growth can be read off Table 8.12 directly. Because the graph is a living artifact (Section 8.3), the Run 1→Run 2 comparison below is simply the first such re-evaluation cycle on record, and each future regeneration will likewise replace the headline figures wholesale and carry the prior build forward for comparison. Both builds carry the same *pre-correction* GT layer (2,439 human-annotated mentions resolving to 852 GT artworks), so the reconciliation to the corrected 2,422-mention corpus (Section 7.4.7) remains outstanding in Run 2 as it was in Run 1.

The direction of change is informative. Materializing the ObliquER artworks lifts the entity layer from 852 to 15,804 artworks and drives the OOKB share from 63.5 % to 79.3 % (12,529 of 15,804). Widening the resolved scope from the stratified ground truth to the full recognition layer recovers a far larger long tail of works that Wikidata does not cover – the canonization signature taken up in Chapter 11, now visible at corpus scale. As noted in Section 8.3.2, the entity counts are an upper bound, since cross-variant deduplication is incomplete (the 3,275 linked artworks collapse onto only 3,070 distinct Wikidata QIDs) and only the new ontology-guided variant's artworks have been clustered so far, so a global reconciliation pass will reduce these figures rather than raise them. Strict provenance coverage (M07) rises from 0.664 to 0.751, not because any annotation lost or gained provenance but because the machine-extracted, paragraph-anchored fraction grew. The residual outside the strict set is all but identical across builds (Run 2: 27,898 annotations = 2,439 human GT mentions + 25,459 non-mention nodes; Run 1: 27,908, the ten-node difference being the self-pair co-occurrence annotations dropped in the co-occurrence-layer correction), while the strict-satisfying count grew from 55,246 to 84,215 (Section 9.5.4).

Table 8.12: Population Run 2 (the canonical build) measured against the prior Run 1. Both columns report measured counts from the respective built graph; the Run 2 column is the source of every KG figure reported in this dissertation, and the Run 1 column is retained only to document the change. Triple counts are over the prompt-stripped serialization (Section 8.3); the GT layer is the pre-correction layer in both builds.

Quantity	Run 1 (prior build)	Run 2 (canonical)
GT mentions (human layer)	2,439	2,439
implicit / explicit	811 / 221	811 / 221
ObliquER recognition mentions	55,246	84,215
Typed mentions (GT + ObliquER)	57,685	86,654
Legacy NER + co-occurrence (untyped)	25,469	25,459
Total annotations (`oa:Annotation`)	83,154	112,113
Provenance activities	6,965	13,794
Artwork entities	852	15,804
Wikidata-linked / OOKB	311 / 541	3,275 / 12,529
OOKB share	63.5 %	79.3 %
Triples per mention (M05)	9.1	9.7
Strict provenance (M07)	0.664	0.751
Total triples (deployed base)	1,671,710	2,448,545
Total triples (M01, closed)	1,673,859	2,455,976

8.4 Website and user interface: final implementation

The interface reported in this section constitutes the deployment half of empirical contribution **E3** of Chapter 1: a web frontend that exposes the populated KG reported in Section 8.3 to non-technical users, demonstrating that hermeneutic provenance (**C3**) is implementable at production scale. The Viewsari web application is accessible at `https://viewsari.ise.fiz-karlsruhe.de` and constitutes the primary access point for the KG constructed in this dissertation. The interface exposes both the gold GT corpus (16 biographies, 232 annotated paragraphs across 6 volumes) and the broader populated KG (165 biographies and 3,479 paragraphs across all 10 volumes, with 86,654 typed mention annotations from the canonical Run 2 build).

The deployed views follow the architecture documented in Section 7.7; the application opens on a landing page (see Figure 8.4) that carries a hero statement, headline statistic groups, a featured bookshelf of the 16 GT biographies, and short teasers linking to the main components reachable from the top navigation bar. From the landing page, two reading paths diverge. The biography paragraph viewer (Section 7.7.2) is the central reading interface, with a split-pane layout juxtaposing the de Vere facsimile against annotated paragraph text; mentions are color-coded by

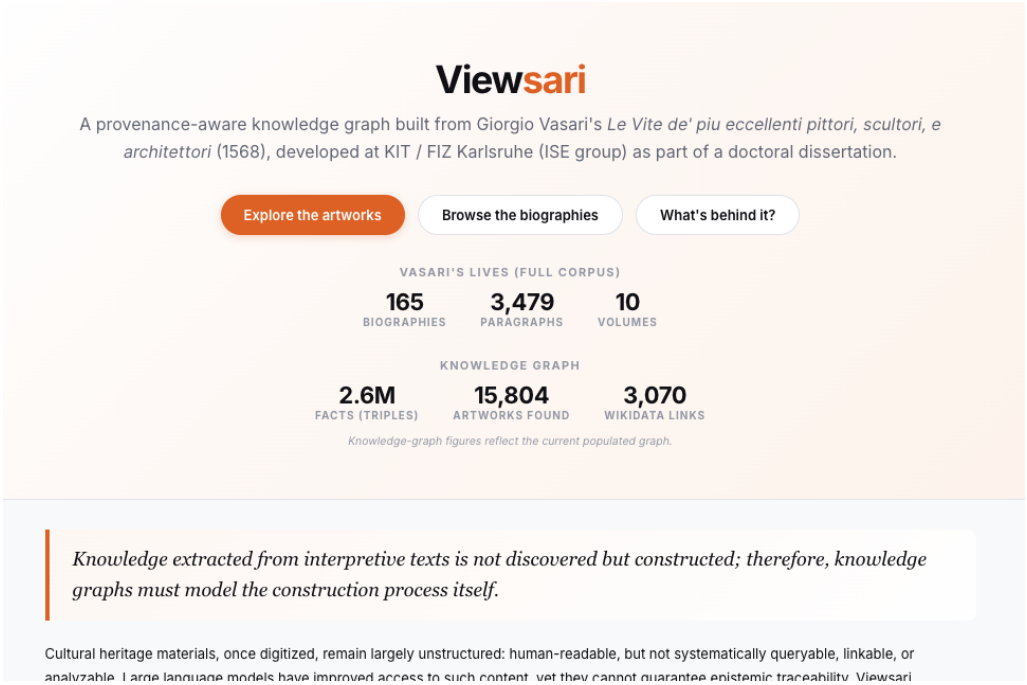


Figure 8.4: Landing page of the Viewsari web application, presenting corpus and KG statistics after Population Run 2, alongside the system's guiding design principle that knowledge extracted from interpretive texts is constructed rather than discovered, and must therefore model its own construction process.

type (amber for explicit, blue for implicit, purple for coreferent, gray for generic), and clicking a mention opens a detail panel with metadata, entity identifier, and Wikidata integration. The knowledge-base browser (Section 7.7.3) is the corresponding entity-focused path through which users can browse and explore the knowledge base's entities, with a new update that includes a faceted search (see Figure 8.5).

Two further views support exploratory navigation rather than sequential reading. The KG explorer (Section 7.7.4) renders interactive force-directed views (D3.js) over pre-computed subgraphs, with nodes typed and color-coded – person in pink/magenta, artwork in amber/gold, co-occurrence in teal, biography in purple – and Wikidata-fetched images. The corpus-level annotations dashboard, the deployment-time addition, provides a tabular overview of extraction outputs per biography with aggregates over annotated paragraphs, mention counts by type, unique entities, and the Wikidata linking rate. For users who require direct access to the underlying data, the SPARQL query interface (Section 7.7.5) is a browser-based query editor with syntax highlighting, tabular result display, and pre-defined example queries covering representative competency questions, while the ontology documentation is mounted as a static portal generated from the OWL source

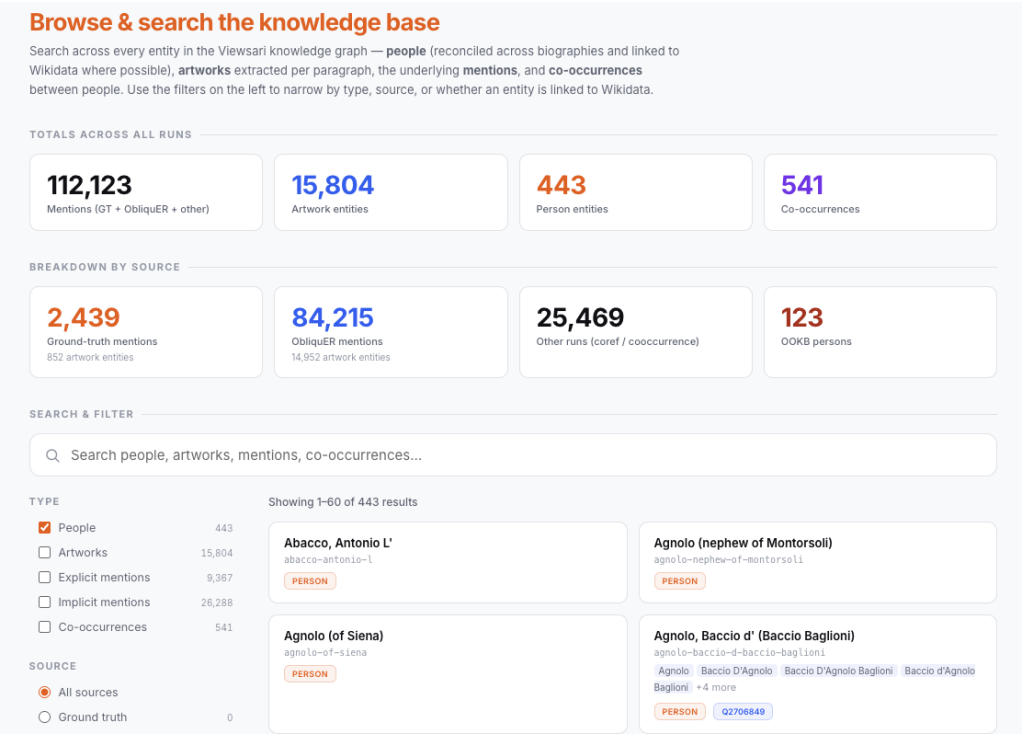


Figure 8.5: The Viewsari knowledge base browser, showing aggregate statistics across all extraction runs (112,113 mentions, 15,804 artwork entities, 443 person entities, 632 co-occurrences) with a breakdown by source distinguishing GT annotations, ObliquER output, and other runs, alongside a filterable entity search interface.

by WIDOCO. Finally, the ObliquER pipeline demonstrator (Section 7.7.6) is an interactive walkthrough of a recognition pass on a representative paragraph, showing the prompt, the response, and the resulting mention annotations side by side.

Functional requirements covered by the interface. Table 8.13 maps the functional requirements from Phase I (Section 7.3.1) and Phase II (Section 7.3.2) requirements engineering to the interface views that address them.

All nine sampled requirements are addressed. The inline provenance visualization (RQII.30) is the most novel interface feature compared with the comparison systems. It renders the full `oa:Annotation/prov:Activity` chain for each selected mention as an interactive graph, making the extraction provenance inspectable without any knowledge of SPARQL or RDF.

The three design principles that govern the interface – ontological transparency, provenance visibility, and progressive disclosure – are stated and motivated in Section 7.7.7 of the Methodology

Table 8.13: Functional requirements coverage by Viewsari interface view.

Requirement	Description (abbreviated)	View
RQI.8	Index search for extracted resources	Annotations, KG Explorer
RQI.9	Semantic and visual search	KG Explorer, SPARQL
RQI.10	Open-ended browsing	KG Explorer (random)
RQI.11	Visibility of connection strength	KG Explorer (edge labels)
RQI.12	Visualization of resources	KG Explorer, Annotations
RQII.30	Inline provenance per annotation	Annotation viewer drawer
RQII.31	Links to source text passages	Facsimile pane, page anchors
RQII.32	Ontology documentation access	Ontology, Ontology Docs
RQII.33	SPARQL query interface	SPARQL

chapter. The deployed implementation realizes each of them. The bibliographic breadcrumb mirrors `frbr:isPartOf`, mention color coding follows the `viewsari:mention` type hierarchy, the OOKB visual marker distinguishes externally resolved entities from Viewsari-internal IRIs, and the layered views, from the homepage through the SPARQL editor, span the spectrum of analytic granularity.

9 Evaluation

This chapter evaluates this dissertation's contributions against formal criteria and comparative baselines.[1] Because the primary contribution is an ontology-engineering and methodological artifact, the evaluation is anchored in the established criteria for such artifacts – logical soundness, competency-question coverage, and epistemic adequacy for interpretive sources (the structural/-functional/usability triad of Gangemi et al. (2006), Section 9.1.1). The overall framework of this dissertation is not measured by ObliquER's extraction metrics; these figures are only one part of the evaluation, capturing how accurately the pipeline recognizes and links mentions rather than the value of the ontology and methodology as a whole. The raw outputs from which these evaluations are drawn are reported in Chapter 8.

The evaluation is organized in three parts: (1) formal ontology evaluation along three dimensions, together with a comparative modeling study against CIDOC CRM (Doerr 2003) and a qualitative assessment of epistemic adequacy; (2) ObliquER performance relative to baselines and in ablation; and (3) an interpretation of the KG population results as methodological claims. The web interface that exposes the populated KG (Section 8.4) is deliberately omitted from this evaluation. Interface design and web application development are not a research focus of the present work, and the website is treated as a means to an end – the access surface through which the ontology and KG become explorable – rather than as a contribution whose qualities require systematic assessment in their own right.

9.1 Ontology evaluation

This section presents the formal ontology evaluation along three dimensions and describes the outcomes, thereby building the baseline for the interpretation and discussion in Section 11.2.

[1] The ObliquER baseline comparison, ablation, and cross-model evaluation reported in Section 9.4 appear in preliminary form in the ObliquER paper (under review, CIKM 2026) and are substantially extended here.

9.1.1 Evaluation design

For the evaluation of the Viewsari ontology, we assess logical soundness, empirical effectiveness in KE, and epistemic adequacy for interpretive sources. This multi-dimensional framing follows the structural/functional/usability triad of Gangemi et al. (2006), adapted here so that the third axis explicitly addresses epistemic adequacy for interpretive sources rather than usability profiling in the general sense, reflecting the dual nature of the ontology as both a formal OWL artifact (subject to automated verification) and a scholarly knowledge representation (subject to domain validation and qualitative assessment). Concretely, the logical-consistency check of Section 9.1.2 instantiates the *structural* dimension (graph-level logical-adequacy properties such as consistency and class satisfiability), the competency-question coverage of Section 9.1.3 instantiates the *functional* dimension (precision- and recall-style fitness against the ontology's intended use), and the qualitative inspection of Section 9.3 re-scopes the *usability-profiling* dimension from interface-pragmatic annotations to the epistemic bindings that the humanities setting elevates to first-class evaluation criteria. Results from the three dimensions are reported in Section 9.1.2–9.3.

9.1.2 Logical consistency

This subsection instantiates the *structural* dimension of the Gangemi et al. (2006) definition. It assesses logical-adequacy properties of the ontology graph (class satisfiability and global consistency) under context-free, syntax-level criteria. Logical consistency of the Viewsari ontology is verified using the HermiT OWL reasoner (Glimm et al. 2014) applied to the Phase II ontology in its current form (Section 7.3.2). Because HermiT does not scale to the full $\sim$2.45M-triple populated graph in practical time, the reasoner-level evaluation is restricted to the synthetic ABox described in Section 7.6.4 (27 individuals, 323 triples, with at least one instance per class and one usage per object property), loaded with the TBox into Protégé 5.6.7 and classified with HermiT 1.4.3 after removing the FaBiO (Peroni and Shotton 2018) and PROV-O (Lebo et al. 2013) import chain.

HermiT classifies the combined TBox and synthetic ABox as consistent, which means that all classes are satisfiable, no individual is assigned to both a class and its complement under the asserted axioms, and the ontology conforms to the OWL 2 DL profile. No additional reasoner output is presented, as the dissertation does not address the realization of specific inferred subclass axioms or the computational cost of classification on the synthetic ABox. The primary claim is the consistency result; downstream users requiring the full entailment closure may independently apply HermiT to the published TBox. For the scaling reason noted above, full-graph behavior is assessed instead via competency-question coverage (Section 9.1.3) rather than description-logic reasoning.

Logical consistency is a necessary but not sufficient condition for ontological quality. The reasoner's verdict establishes that the formal axioms do not contradict one another, but does not assess whether they correctly represent the domain. The formal adequacy of the ontology for the Vasari domain is assessed through CQ coverage (Section 9.1.3) and the comparative modeling study (Section 9.2). The interpretive reading of this judgment against the four contributions C1–C4 – in particular about the three-layer architecture and the deliberate scoping of the reasoner check itself – is taken up in Section 11.2.

9.1.3 Competency question coverage

Following Gangemi et al. (2006)'s functional dimension, we operationalize ontology fitness as the answerability of CQs on the populated graph. We distinguish three outcomes prior to reporting results:

- *fully answerable* – the result set is non-empty and structurally consistent with the expected answer type;
- *partially answerable* – the query returns a structurally correct but empty result set, traceable to a population (data) gap or a deliberately deferred modeling extension (a distinction established in competency question methodology by Grüninger and Fox (1995)); or
- *unanswerable* – the query fails to parse or raises a runtime error, which would indicate a hard modeling gap.

A residual subset of 18 catalog entries comprises *descriptive* questions concerning data reuse, teaching, open-ended exploration, the functionality of the web interface, or federated querying. These are reported separately and not counted against SPARQL coverage; they are not thereby left unmet, however, but discharged by the Viewsari web application and its documentation rather than by SPARQL – the KG explorer, the SPARQL editor and endpoint, and the WIDOCO ontology portal each answer a subset, with the per-question mapping (and the facets deferred pending unpopulated entity types) given in Section 7.7.

The 70 CQs together with their SPARQL translations are maintained as a single catalog (Section A) and executed against the closed populated graph reported in Section 8.3; reproducibility, evaluation scripts, and serialization choices are described once in the project repository and not repeated here.

Closure of the per-instance *in-paragraph* edge Five Phase-I queries (CQI.3, CQI.12, CQI.19, CQI.21, CQI.22) require the *in-paragraph* edge on cooccurrence and person nodes. The edge is currently given in the closed graph, our extension of the deployed graph that aligns the data with the ontology's intended traversal pattern (*extracted content* → *in paragraph* → *paragraph*) (cf. Section 8.3).

Aggregate coverage Executed on the closed graph, 39 of the 52 SPARQL-translatable CQs are *fully answerable* (75.0 % of the SPARQL-translatable subset, 55.7 % of the full catalog) and 13 are *partially answerable* – the query parses and is structurally correct but returns an empty result set traceable to a known population gap rather than to a modeling defect. No CQ is *unanswerable*, and every SPARQL-translatable query parses and executes without runtime error and either binds or signals an empty result that the catalog flags in advance. The remaining 18 CQs are descriptive. Table 9.1 consolidates these figures. The 13 partial cases trace to a small set of population gaps, each a matter of deferred scope rather than a modeling defect. The *location* and *historical-event* classes are declared in the ontology but carry no instances yet, since the present work is scoped to artworks; typed artist–artist relations such as `viewsari:hasPupil` are not extracted, as the current pipeline produces only person–person co-occurrences; the co-occurrence weighting score is not yet materialized; artwork participation in co-occurrences is unsupported for the same person-only restriction; and the Italian-edition cross-manifestation requested by the domain experts has not yet been ingested. These results are discussed cluster-by-cluster after Table 9.3.

The headline figure here is not the 39 fully answerable questions but the *zero* unanswerable ones. Every question a historian or engineer brings to the ontology can be put to the graph, and where the graph is silent (the 13 partials), it names exactly which data is missing rather than failing to ask. This is the point at which the dissertation's opening claim is redeemed in concrete terms.

Table 9.1: CQ coverage of the Viewsari ontology (closed KG with 2,455,976 triples). Percentages in the SPARQL-only column are calculated across the 52 SPARQL-translatable CQs; percentages of the total are calculated across all 70 catalog entries.

	Count	% of 70	% of 52
Fully answerable	39	55.7	75.0
Partially answerable	13	18.6	25.0
Unanswerable	0	0.0	0.0
Non-SPARQL (descriptive)	18	25.7	–
Total	70	100.0	100.0

The query-complexity distribution (Table 9.2) shows that the catalog stresses the ontology's *traversal* structure rather than its aggregation surface. Given our focus on representing complex provenance chains, a majority of CQs require multi-hop graph patterns spanning at least two of the three representational layers (bibliographic, structural, and content). Every multi-hop CQ parses and executes; the ten multi-hop CQs that remain partial – CQI.15–18, CQII.3, CQII.9, CQII.11–12, CQII.16, CQII.24 – return empty result sets that resolve to population rather than traversal failures.

Table 9.2: SPARQL query complexity distribution.

Complexity	Count
multi-hop (two or more joins)	28
simple (single triple pattern + filter)	15
aggregation (`COUNT`/`GROUP BY`)	9
Sum (SPARQL-translatable CQs)	52

The per-cluster breakdown in Table 9.3 shows that the 13 partial cases concentrate in the clusters whose questions depend on entity types that have been declared in the ontology but not yet populated with instances. The *Locations* cluster (3 of 4 SPARQL-translatable CQs partial) and the location-dependent rows of *Artworks, Locations, and Entities* (4 of 5 partial) all probe `viewsari:0001015` (location), of which the deployed graph contains zero instances; the *Historical Events* cluster (2 of 2 SPARQL-translatable CQs partial) probes `viewsari:HistoricalEvent`, likewise unpopulated. The remaining four partials are distributed across other clusters. CQI.7 (scores on cooccurrences, not yet materialized), CQII.3 (typed artist–artist relations such as `viewsari:hasPupil`, deferred to a future iteration), CQII.9 (artwork participation in cooccurrences, which the current pipeline restricts to person–person), and CQII.24 (Italian-edition manifestation, absent from the current KG). The clusters that retain descriptive (non-SPARQL) CQs – *User Interaction and Exploration*, *Search and Data Reuse*, *Linked Data Enrichment* – correspond to research-design questions whose answers are documented in the project repository.

Further interpretive reading of these results against C1–C4 and RQ1 is taken up in Section 11.2.

Table 9.3: CQ coverage per thematic cluster (post-closure).

Cluster	Full	Partial	Unans.	Descr.	Total
Artists and Co-occurrences (Phase I)	7	1	0	0	8
Artists, Collaborations, Relationships (Phase II)	3	1	0	1	5
Artworks (Phase II)	3	1	0	1	5
Artworks, Locations, Entities (Phase I)	1	4	0	0	5
Bibliographic and Provenance (Phase I)	5	0	0	0	5
Evidence and Bibliographic Information (Phase II)	4	1	0	1	6
Extraction Provenance and Epistemic Modeling	7	0	0	1	8
Historical Events (Phase II)	0	2	0	2	4
Linked Data Enrichment	2	0	0	2	4
Locations (Phase II)	0	3	0	1	4
Persons and Attributes (Phase I)	5	0	0	0	5
Search and Data Reuse (Phase I)	2	0	0	4	6
User Interaction and Exploration	0	0	0	5	5
Total	39	13	0	18	70

9.2 Comparative modeling study: Viewsari vs CIDOC CRM

The comparative modeling study examines what can and cannot be represented when the same source passages from Vasari's *The Lives* are modeled using parts of the CIDOC CRM family.

We begin with a comparative study of vanilla CIDOC CRM, CIDOC CRM combined with PROV-O, and the Viewsari model. We broaden this comparison to include CIDOC CRM extensions that are closer to our goal, namely CRMdig and CRMinf. The comparison is conducted as a *requirements-based fit-for-purpose* evaluation (Vrandečić 2009, Suárez-Figueroa et al. 2015, Blomqvist et al. 2016), in which each candidate ontology is assessed against a set of project-specific representational requirements, in our case, derived from the analysis of Vasari's text and the conceptual contributions developed in Section 7.3.1 and Section 7.3.2. This evaluation method follows the standard approach in ontology engineering for justifying the development of a new ontology against existing alternatives. We thereby assess the representational requirements that motivate the new ontology in the first place (Hlomani and Stacey 2014, Keet 2018). We anchor these representational requirements to the following comparisons: (R1) the capacity to anchor entity mentions to character-offset spans in a digital source text; (R2) the capacity to represent the computational particulars of a generative extraction run, including extraction agent,

prompt template, model version, and run-level configuration; and (R3) the capacity to represent mentions of entities without determinable real-world referents (OOKB implicit mentions). The purpose is to show concretely what is representable, what is flattened, what becomes ambiguous, and what is lost in each case, with respect to the named requirements. A full structural alignment between Viewsari and the CIDOC CRM family, or a quantitative benchmark of expressiveness on a shared corpus, lies beyond the scope of this work; given that the candidate ontologies operate at different layers of the modeling problem, such alignment would constitute a category mismatch rather than a methodological gap. The CIDOC CRM family is chosen as the comparison target precisely because it is the de facto standard CH vocabulary and the one structurally closest to Viewsari, with CRMdig the closest existing vocabulary for computational-extraction provenance and CRMinf the dedicated argumentation/inference extension (Section 5.2.2.1), and therefore functions as the *best-case* existing candidate. Because it is the best-case candidate and yet natively expresses none of R1–R3, the negative result generalizes a fortiori to less expressive or more distant vocabularies. The claim is deliberately scoped to native expression and schema-level enforcement. As the combinatorial analysis below shows, a union of existing vocabularies can be made to carry R1–R3, but only by reconstructing the mention/referent inversion that is not yet made explicit or obligatory. An exhaustive cross-vocabulary survey is therefore neither necessary for the representational claim nor the kind of evidence it requires.

9.2.1 The unifying claim: textual mention versus referent

The comparison aims to show that the CIDOC CRM family – core CRM for CH events, CRMdig for digital provenance, CRMinf for argumentation, among others – is organized around the *referent*. A referent is a human-made object, the observable entity. None of these, individually or in combination, provides a mechanism for representing the *textual mention* as a categorically distinct entity from such a referent. A mention, a character span in a specific paragraph of a specific edition, is precisely what our approach surfaces and what the PIEA introduced in Chapter 1 suppresses by treating reference-to-referent as completed prior to representation.

We illustrate the modeling gap between CIDOC CRM and Viewsari by the example sentence “He painted a panel for the friars of Santa Croce, representing the Last Supper” (Life of Ghirlandaio), in which an artwork is named only by its depicted subject and commissioner. Figure 9.1 shows that CIDOC CRM asserts the panel as a real-world object, without further information on its context as stemming from a specific text.

In Figure 9.2, we see that CIDOC CRM + PROV-O can attach an extraction activity to the panel node, but thereby conflates the panel’s existence with the act that identified it (possibly a mention of it somewhere).

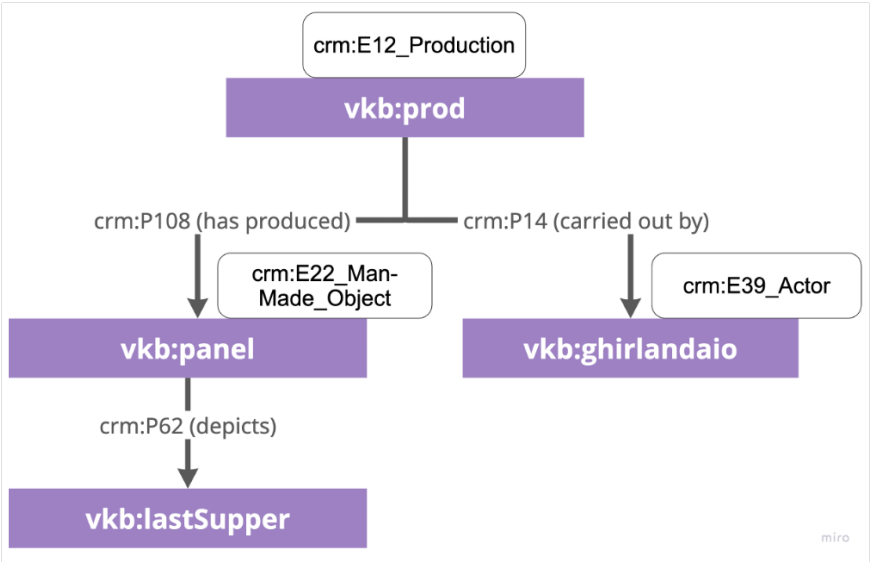


Figure 9.1: Implicit artwork mention modeled in CIDOC CRM only. The panel is typed as `crm:E22_Human-Made_Object`, the production event as `crm:E12_Production`. The model asserts the artwork's existence as a real-world entity but provides no mechanism to record that it was *identified through an implicit reference* in the source text. The interpretive act is invisible.

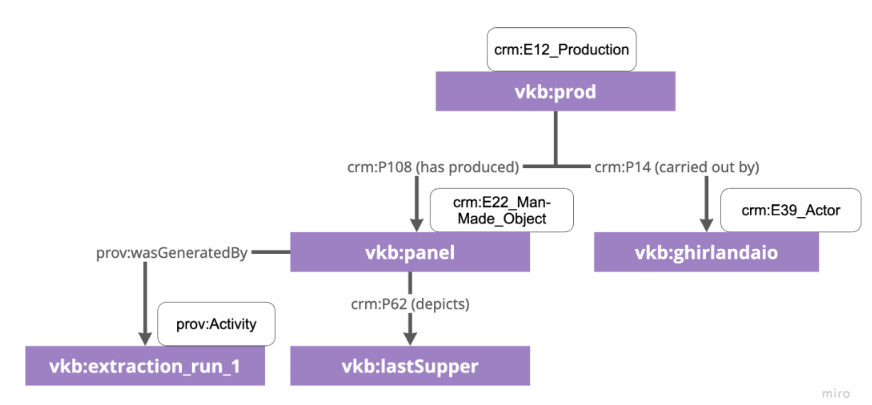


Figure 9.2: Implicit artwork mention modeled in CIDOC CRM combined with PROV-O. Adding `prov:wasGeneratedBy` links the panel entity to an extraction activity, but this is semantically incoherent. The physical panel predates Vasari is not an output of any computational process. PROV-O's `wasGeneratedBy` conflates the *knowledge claim about* the artwork with the artwork itself.

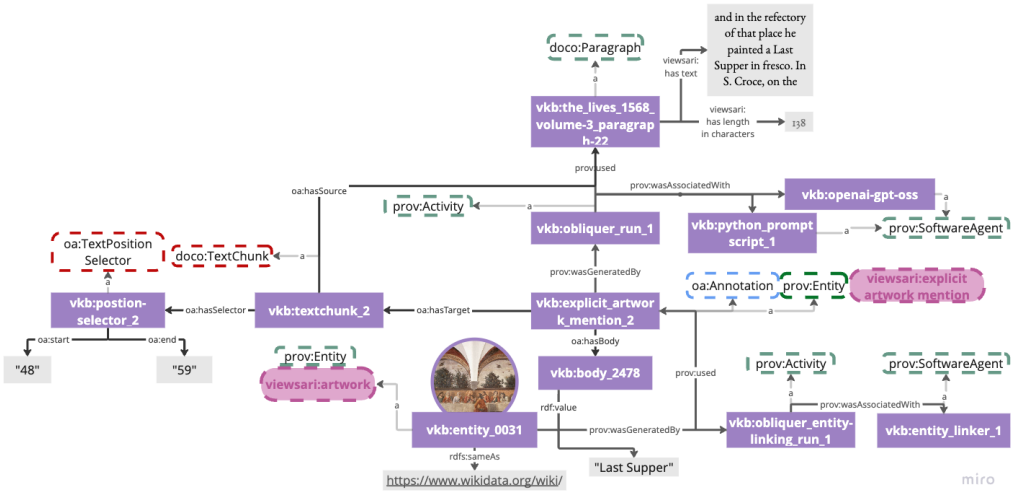


Figure 9.3: Example instantiation of a Viewsari representation of an implicit artwork mention (the 'Last Supper'), showing two of the three representational layers, with a focus on the extraction process and result, as well as the process of linking the mention to a concrete artwork (the referent).

As Figure 9.3 illustrates, Viewsari types the mention itself as an `oa:Annotation` with a character-offset target and a prov-attributed extraction activity, leaving the panel's knowledge-base status to a separately resolved artwork node, the actual referent resolved through an EL process.

When compared in terms of *representational expressivity*, whether a given model can draw the distinctions required by the corpus (e.g., extraction provenance), and *epistemic adequacy*, whether the model not merely permits but structurally requires these distinctions to be recorded such that a representation produced under the model is epistemically accountable by construction, the CIDOC CRM family can, in principle, be extended to express some of these distinctions, but no existing extension makes them obligatory. Viewsari addresses both dimensions simultaneously and produces a more verbose triple structure than CIDOC CRM alone – the design budget of 12–18 triples per mention (derived in Section 9.2.2) is undershot by the 9.7 triples per mention on the closed graph (Section 9.5.1), and the 100 % provenance coverage of Section 9.5.4 reports the population side of the same trade. The interpretive reading of the verbosity–provenance trade against C1 and C4 is taken up in Section 11.2; the detailed co-occurrence modeling and the formal CRMinf scope analysis are taken up in Section 9.2.3 and Section 9.2.4, respectively.

9.2.2 The combinatorial objection and its limits

A natural objection to the preceding analysis runs as follows. Rather than introducing a new ontology, one could simply combine CIDOC CRM, PROV-O, and the Web Annotation Ontology (Sanderson et al. 2017) (`oa:`) to cover the same representational ground. Each vocabulary is mature, widely adopted, and independently interoperable with the CH stack. Why should their union be less expressive than Viewsari?

The objection rests on a category error. The three vocabularies are not inexpressive; they are expressive at the wrong ontological layer. Their combination yields a richer account of CH facts and data lineage, but it does not provide what this dissertation requires. A representation of interpretive claims about the text that is structurally prior to and independent of any claim about the referent's existence. The mismatch is not a matter of degree but of structural layer, as the following makes concrete.

Incoherent subject assignment. PROV-O's `prov:wasGeneratedBy` requires an entity that is the *output* of an activity. When CIDOC CRM + PROV-O is used to represent an extracted artwork mention, the only available subject is the artwork node itself, typed, say, as `crm:E22_HumanMade_Object`. Attaching `prov:wasGeneratedBy` to that node asserts that the physical panel was generated by ObliquER, which is semantically incoherent. The panel, if it exists, predates Vasari by decades and is not the output of any computational process. To avoid this, one may introduce a wrapper node, a knowledge-claim or information object, that is the actual output of the extraction activity, while the artwork remains a separate entity linked to it. That design move is precisely the mention/referent inversion that defines Viewsari. Whoever builds this wrapper has, in effect, reconstructed the core structural commitment of `oa:Annotation` over an `oa:SpecificResource` target, without a schema that makes the commitment explicit or obligatory. This is an inherent part of CIDOC CRM. As a reference ontology for resolved real-world entities, it practically embodies PIEA. With the referent being its presupposed subject, materialized before any claim is recorded, it could only reach the required level of representation by introducing the mention as the actual output of extraction, that is, by reconstructing, against CIDOC CRM's design, the very mention/referent inversion that it presupposes away. The deficit is thus not a missing property but a prior commitment, and it is this commitment, rather than any single absent class, that the comparative study isolates.

Provenance depth and parametric inputs. PROV-O provides `prov:wasGeneratedBy`, `prov:wasAttributedTo`, and `prov:wasDerivedFrom` as its core provenance vocabulary. These are sufficient for data-lineage tracking but lack native support for the parametric inputs of

a generative extraction run. Mention type, prompt template version, model identifier, decoding temperature, or confidence score. The COMP-PROV pattern introduced in this dissertation (Section 10.3.2) is not an alternative arrangement of PROV-O properties; it is a design pattern that promotes these parameters to first-class provenance attributes, enabling systematic comparison across extraction runs and supporting the epistemic accountability claim of C4. None of the CIDOC CRM family, PROV-O, or the Web Annotation Ontology natively provides this pattern. COMP-PROV constitutes a pattern rather than a primitive; the relevant issue is not that PROV-O is incapable of supporting it, but that these vocabularies do not include it, as each omits the explicit commitments required for accountable provenance extraction. In this arrangement, the annotation, the typed textual mention, serves as the provenance-bearing object, not the referent. This distinction preserves the coherence of `prov:wasGeneratedBy`. The extraction run is reified as a first-class provenance object, with mention type, prompt template version, model identifier, and decoding parameters exposed as queryable attributes rather than relegated to free-text notes. COMP-PROV introduces this structure precisely. Its portability derives from its status as an extension of a standard PROV-O chain, not from triviality.

Mention typology. The W3C Web Annotation Ontology provides `oa:Annotation` with a flexible body/target model, but no ontological distinction between an explicit mention, one where the artwork is named directly, and an implicit mention, one where it must be inferred from contextual description. This is not a missing property; it is a missing concept. The four-part typology (explicit, implicit, coreferent, generic) introduced in Viewsari (Section 7.4.6.1) defines categorically different resolution strategies. Implicit mentions require contextual inference, coreferent mentions require coreference chains, and generic mentions are deliberately excluded from EL. A plain `oa:Annotation` body cannot distinguish these cases without an external vocabulary that supplies the typing, and no such vocabulary exists in the CIDOC CRM family or in PROV-O. The typology is an original contribution of Viewsari, not a derivable consequence of stacking existing schemas.

Beyond distinguishing resolution strategies, the typology is also an implicit formulation of certainty and uncertainty about the referent, and in this, it is a source-critical device that CIDOC CRM lacks. The difference between an explicitly named mention and one recoverable only through contextual inference is, in source-critical terms, a difference in the certainty with which the source text commits to the referent. An explicit mention is a high-certainty textual witness, whereas an implicit one is a lower-certainty inference that the reader, or the extraction model, must underwrite, and a generic mention deliberately withholds referential commitment altogether. The typology therefore encodes graded certainty about the referent directly in the annotation's structure, prior to and independent of any entity resolution. CIDOC CRM has no place for this gradient. Its existential commitment either admits a referent or it does not, and even CRMinf

locates belief and doubt at the level of propositions *about* world entities (Section 9.2.4) rather than in the manner in which a text *mentions* them. Viewsari's typology thus carries a source-critical reading of certainty that the CIDOC family cannot express without, once again, reconstructing the mention/referent inversion.

The reconstruction also determines the verbosity budget evaluated in Section 9.5.1. When modeled exclusively in CIDOC CRM, as shown in Figure 9.1, the implicit panel mention requires six triples, but this approach omits the mention itself, its character offset, and the extraction act. In contrast, the Viewsari scaffold captures all of these elements. This scaffold, as depicted in Figure 9.3, can be enumerated directly from the figure. The irreducible per-mention core consists of the `oa:Annotation` with its source, target, body, and `prov:wasGeneratedBy` edge; the `doco:TextChunk` target with its `oa:TextPositionSelector` and two offset literals; and the typed textual body. This configuration requires 12 triples per mention, including those not present in the knowledge base. When a mention is resolved to a referent, the model adds entity-linking provenance (`prov:wasGeneratedBy` and `prov:used` on the body), the typed referent node, and the `rdfs:sameAs` edge to Wikidata, increasing the total to approximately eighteen triples. The 12—18 triple range is not an arbitrary estimate but reflects the difference between an unlinked and a fully resolved mention. A combinatorial reconstruction using CIDOC CRM, PROV-O, and the Web Annotation Ontology elements yields the same range, as it must reconstruct the identical scaffold; it does not reduce the triple count, but omits the schema that enforces the scaffold. The shared paragraph and extraction-run nodes are amortized across all mentions within a paragraph. This, combined with generic and out-of-knowledge-base mentions that lack a linking chain, accounts for the observed mean of Section 9.5.1 (9.7), which falls below the twelve-triple minimum.

The objection therefore treats the difference as quantitative, a matter of missing properties or underspecified classes that further extension could supply. The asymmetries above show that the difference is qualitative. CIDOC CRM operates at the layer of *cultural heritage facts*: it models what is known about real-world objects, events, and actors. PROV-O operates at the *data* layer, modeling how data objects were derived from other data objects. Viewsari uses this to operate on the layer of *interpretive claims about text*. It models what a computational process inferred from a specific character span in a specific textual source, under specific parametric conditions, without presupposing that the inference refers to an existing entity. Combining the two established layers does not produce this one.

Table 9.4 summarizes the comparison across the dimensions developed in Section 9.2.1 and Section 9.2.3.

Table 9.4: Comparative representational capabilities. CIDOC CRM, the CIDOC CRM + PROV-O + Web Annotation stack, and Viewsari across six dimensions relevant to generative, provenance-aware extraction from interpretive humanities texts.

Dimension	**CIDOC CRM**	**CIDOC + PROV-O (+ OA)**	**Viewsari**
Primary subject of representation	Referent (`E1_CRM_Entity`)	Referent + data provenance wrapper	Mention (`oa:Annotation`)
Existential presupposition	Yes: every instance presupposes the entity exists or has existed	Yes: retained on the artwork node; wrapper does not resolve the referent's status	No. The mention asserts only that a character span exists in the text
OOKB entities	Not representable without existential commitment	Not representable without existential commitment	First-class IRI anchored to mention; referent resolution is optional
Mention typology	Not present	Not present	Explicit / implicit / coreferent / generic, each with distinct resolution semantics and an implicit certainty gradient
Extraction-run parameters as provenance	Not present	Partial via `prov:Activity`; no native class for prompt template, model ID, or confidence score	COMP-PROV pattern: parametric inputs are first-class provenance attributes
Interoperability with referent-based KBs	Native (designed for this purpose)	Native	Via EL (downstream, optional)

9.2.3 Co-occurrence modeling and the difference of ‘interpretations’

We further illustrate in Figure 9.4 the dimensions on which the Viewsari model differs from existing CIDOC CRM extensions when the referent and the mention are not the same, and when ‘interpretation’ means something different in each case. For this, we use the textual co-occurrence of the two artists Cimabue and Giotto from the text: “Cimabue, going one day from Florence to Vespignano, found Giotto, who, while his sheep were feeding, was drawing one of them on a stone. . . ”.

As can be seen, the output node has no semantic type, no character-offset anchor within the source paragraph, and no record of the epistemic character of the extraction act. The gap is structural. CRMdig’s provenance chain runs from the physical object to its digital representation, not from digital text to an interpretive knowledge claim.

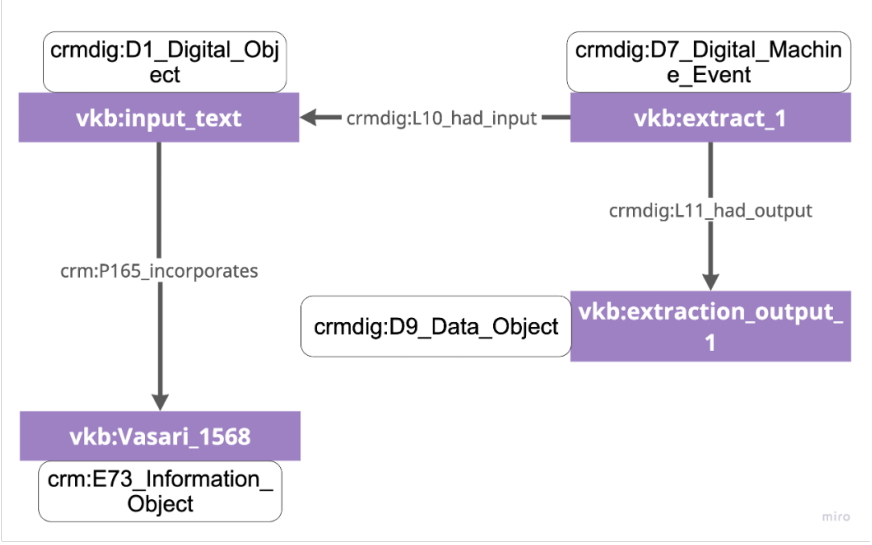


Figure 9.4: Co-occurrence extraction modeled in CIDOC CRMdig. The extraction run is typed as a `crmdig:D7_Digital_Machine_Event` whose input is a `crmdig:D1_Digital_Object` incorporating the Vasari 1568 edition (`crm:P165_incorporates`), and whose output is a `crmdig:D9_Data_Object`.

9.2.4 The CRMinf scope boundary

Figure 9.4 shows the modeling approach with CRMdig. This CIDOC extension is the closest analog for representing extraction-derived data: it provides classes such as `crmdig:D7_Digital_`

`Machine_Event` for documenting digital activities, `crmdig:D1_Digital_Object` for the digital text fed into the activity, and `crmdig:D9_Data_Object` for its output (Figure 9.4). Under this model, the Cimabue and Giotto co-occurrence would be represented indirectly as the output (`crmdig:L11_had_output`) of a digital event whose input (`crmdig:L10_had_input`) is a digital text, with the digital text linked via `crm:P165_incorporates` to the abstract work modeled as a `crm:E73_Information_Object`. However, no native mechanism exists to anchor the co-occurrence to a specific character offset within the source paragraph (Tzitzikas et al. 2022); the resulting `D9_Data_Object` has no semantic type for the extracted content (co-occurrence, entity mention, relation), and the co-occurrence itself remains a relation between entities rather than a substantive object that carries metadata, weight, or provenance.[2]

This limitation is structural. CRMdig was designed for scientific observation workflows – digitization, satellite data – in which a provenance chain runs from a *physical object* through a measurement event to a digital representation (Doerr et al. 2016).[3] Its central abstraction is `D2_Digitization_Process`. A physical thing is measured and a digital object results. In the Viewsari scenario, however, no physical object is being measured. The input is already a digital text; the output is a semantic interpretation of that text. `D3_Formal_Derivation`, the closest CRMdig class for a computational transformation, presupposes a deterministic, subject-preserving algorithm. A computational method that performs co-occurrence extraction or an LLM that extracts artwork mentions is none of these, since they are probabilistic, context-sensitive, or produce an *interpretation* under specific circumstances rather than a transformation. The structural mismatch is therefore not a gap to be filled by an additional property, but rather a difference in epistemic register between a model built for digitization and one needed for generative interpretation.

From CRMdig to CRMinf. The progression from CRMdig to CIDOC CRM's extension for propositions and beliefs comes naturally (Doerr et al. 2023). It is the closest available analog to the knowledge-producing processes of modeling.

[2] A further CRM family extension, CRMtex (Felicetti and Murano 2021, Doerr et al.), models textual genesis, so the creation of textual entities. We do not adopt it here, given its focus on materiality and semiotic stratification of the inscription rather than the mention-versus-referent distinction. CRMtex types the written text and the linguistic object it carries, but it provides no class that holds a textual *mention* apart from the real-world *referent* it may denote, and no vocabulary for the parametric conditions of a machine extraction. The negative native-expressivity result of Section 9.2 therefore extends to it a fortiori.

[3] Similarly to the CRMsci extension, the Scientific Observation Model, which we did not incorporate here on purpose, because it addresses measurements and observations of mostly empirical experiments: `https://cidoc-crm.org/crmsci`.

CRMinf's surface vocabulary – argumentation, belief, inference, evidence – looks, at first glance, compatible with the claim that LLM-mediated extraction is itself an interpretive act. Its specification, however, defines a scope that explicitly excludes this extraction case. Concretely, the scope statement and the class definitions of CRMinf separate the model from the LLM-extraction case along four axes, each elaborated in the paragraphs below and each visible in the Bologna walkthrough that follows:

(1) **Subject of representation (referent versus mention).** CRMinf encodes propositions and beliefs *about* real-world referents, whereas an ObliquER annotation is a claim *about a text span*; the textual mention has no class anywhere in the CRMinf chain (Section 9.2.1).

(2) **Agent.** CRMinf attributes interpretive activity to a human `E39 Actor`, on the explicit premise that "knowledge resides in humans" (Doerr et al. 2023); a parametrized LLM is not an `E39 Actor` but a `prov:SoftwareAgent`.

(3) **Epistemic attitude (belief and honesty).** `I1 Argumentation` is defined as *honest* inference whose actor *justifies and believes* the resulting `I2 Belief`; these are dispositional attitudes that a generative model cannot instantiate, which is why the specification recommends the `E65 Creation` fallback wherever the honesty of an actor's expressions cannot be assumed.

(4) **Provenance of the generative run.** Neither CRMinf nor the `E65 Creation` fallback provides vocabulary for the parametric inputs of a generative extraction (prompt template, model identifier, run configuration), which the COMP-PROV pattern (Section 10.3.2) instead promotes to first-class provenance.

Difference in commitment. The two models are not mutually exclusive – a CRMdig representation can coexist with a Viewsari representation – but they encode different commitments about what is worth recording. CRMdig documents that a digital event produced an output. Viewsari documents what was extracted, between which entities, anchored to which paragraph, in which biography, in which volume, in which translation, of which 1568 edition, through which extraction activity. The verbosity is the price of full traceability through the interpretive chain. Even CRMinf, the CIDOC extension for argumentation and belief, cannot bridge this gap. It models beliefs *about* world entities rather than annotations *on* text spans, and has no vocabulary for the computational parameters of a generative extraction run (Section 9.2.4). This means, as said before, CRMinf's focus is on real-world referents.

Figure 9.5 reproduces Figure 5 of the CRMinf specification (Doerr et al. 2023). The argument structure encoded there follows a person, Francesca Bologna, who adopts the belief held by Tacitus about Nero's location, producing an `I12 Adopted Belief` linked via `J7 is based`

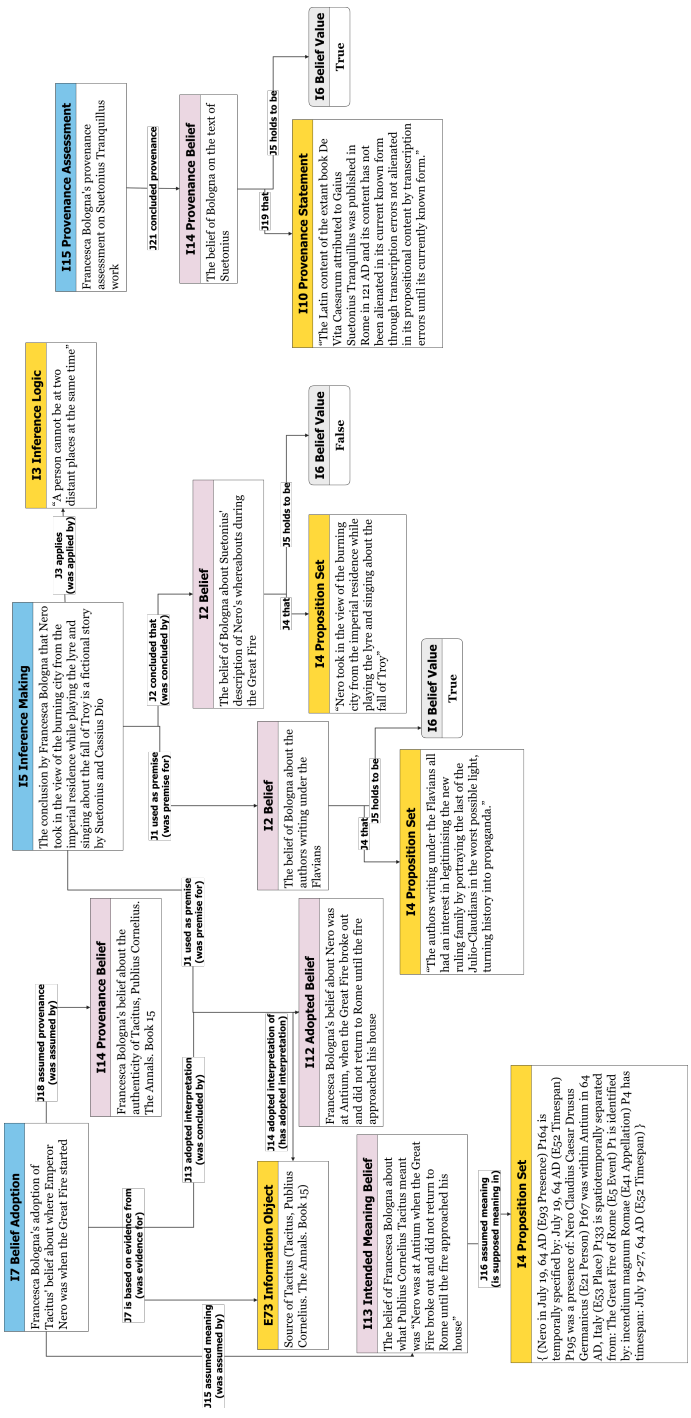


Figure 9.5: Reproduction of CRMinf specification example *A scholarly argument concerning Nero's whereabouts during the Great Fire of Rome* (Doerr et al. 2023), included under the CRMinf CC-BY 4.0 license. Francesca Bologna's argument is encoded as an `I1 Argumentation` decomposed into `I7 Belief Adoption` (from Tacitus), `I5 Inference Making` (rejecting the Suetonius–Cassius Dio account on plausibility grounds), and `I15 Provenance Assessment` (on the textual transmission of Suetonius). The entire chain is carried out by a single human `E39 Actor`.

on evidence from to an E73 Information Object (the surviving copy of Tacitus's *Annals*). Based on premises, Bologna concludes through I5 Inference Making that an alternative story to her known canon is fictional, on the basis of an I3 Inference Logic that "a person cannot be at two distant places at the same time". The CRMinf encoding documents who reasoned over which sources, with which inference logic, leading to which belief states and belief values, and it does so in a form that is interoperable with the rest of the CIDOC CRM heritage stack. The classes I7 Belief Adoption, I5 Inference Making, and I15 Provenance Assessment are precisely the right vocabulary for a scholar reasoning over textual sources.

We extend this argumentation with an example. The class I5 Inference Making makes this explicit in the disclaimer in its scope note: "The class is not intended to promote the *use of computationally decidable systems of logic as replacements of scholarly justifications of arguments*" (Doerr et al. 2023). The intent of CRMinf is to *document* arguments made by scholars, not to *replace* or *produce* them computationally. The LLM-extraction case sits on the opposite side of this disclaimer: it produces structured propositions about the source text from a generative process, and what it requires from an ontology is precisely the documentation of how it produced them, not a framework for inferring whether their content is true. The asymmetry is therefore not between two competing models of the same activity, but between two activities that the two models are respectively designed to document. It builds on the concept that "knowledge resides in humans, who relate the symbols in information to states of affairs in current or past reality" (Doerr et al. 2023).

The model's central instrument, the I2 Belief class, is constructed accordingly as the notion that the associated I4 Proposition Set is to have a particular I6 Belief Value by a particular E39 Actor. A computational system is not an E39 Actor in the CRMinf sense; treating it as one would violate the model's own scope. This can also be transferred to the I1 Argumentation class, CRMinf's general activity for inference, observation, and belief adoption. It is defined as the activity of making *honest* inferences or observations. An honest inference or observation is one in which the E39 Actor carrying out the I1 Argumentation *justifies and believes* that the I6 Belief Value associated with the resulting I2 Belief about the I4 Proposition Set is the correct value at the time that the activity was undertaken and that any I3 Inference Logic or methodology was correctly applied (Doerr et al. 2023). *Justifies* and *believes* are dispositional attitudes, not behaviors that an LLM can instantiate. CRMinf's own authors anticipate the limitation explicitly: "In cases where the underlying honesty of expressions by some actors cannot be assumed, it is recommended to document their statements more generally by instances of E65 Creation, which applies to all propositions, and not as instances of I2 Belief. Otherwise, using CRMinf for documenting the progress of a serious scientific discourse would be compromised" (Doerr et al. 2023). The fallback to E65 Creation is not a Viewsari-equivalent: E65 Creation models

the production of a conceptual object and has no class for the resulting annotation, no edge to a character offset, and no provenance for the activity's parametric inputs.

The contrast for the present co-occurrence example is laid out in Figure 9.6 (CRMinf encoding) and Figure 9.7 (Viewsari encoding).

A fuller CRMinf encoding widens, rather than closes, the gap. The transfer above used the simplified "Vasari enthusiast reads Vasari and adopts the belief" pattern to achieve parallelism with Bologna. A scholarly-fidelity CRMinf encoding of the same claim (the claim that Cimabue and Giotto met) would in fact be considerably richer, and the richness only sharpens the conclusion. The Vasari enthusiast does not simply adopt the meeting as a belief reported in Vasari. He carries out an `I5 Inference Making` activity whose premises are several `I12 Adopted Beliefs`, each obtained through its own `I7 Belief Adoption` from a different source (Vasari's own narration, other early biographers, his own previous publications, modern art-historical literature on Cimabue's late workshop, and chronological evidence from documented commissions). Each adopted belief carries a `J7 is based on evidence from` edge to the `E73 Information Object` that supplied it; each is concluded by its own `I7` activity attributed via `P14 carried out by` to a Vasari enthusiast as an `E21 Person`. The `I3 Inference Logic` that the Vasari enthusiast applies (a plausibility-and-chronological-fit principle, e.g. "a junior Florentine painter active in the 1280s would plausibly have encountered the leading practitioner of the period given documented overlapping commissions") is itself an `E89 Propositional Object` with its own provenance. The resulting `I2 Belief` ("Cimabue and Giotto met"), held by the Vasari enthusiast with `I6 Belief Value true`, is then the conclusion of this whole synthesis. The encoding is methodologically apt for what it models. A scholarly editor's documented argument for a contested historical claim. But it is no closer to what ObliquER does than the simplified version was; if anything, every additional layer of belief, premise, and inference logic anchors the encoding more deeply in the propositional/belief-state layer. The `oa:Annotation` that ObliquER would emit on the same passage – a typed mention of a co-occurrence anchored to characters 142–168 of paragraph 12 of the de Vere translation – has no class in this entire chain, and no place to live as a first-class entity. The more elaborately we encode the Vasari enthusiast's reasoning in CRMinf, the more evident it becomes that ObliquER operates at a different representational layer than CRMinf.

Observation from the walkthrough. The Bologna–Nero example from CRMinf's specification is the most textually oriented canonical example from this standard (Doerr et al. 2023); its proximity to the Viewsari setting makes it the strongest available test of the scope-overlap question. The walkthrough shows that even in this best-case scenario, the four-axis incompatibility persists.

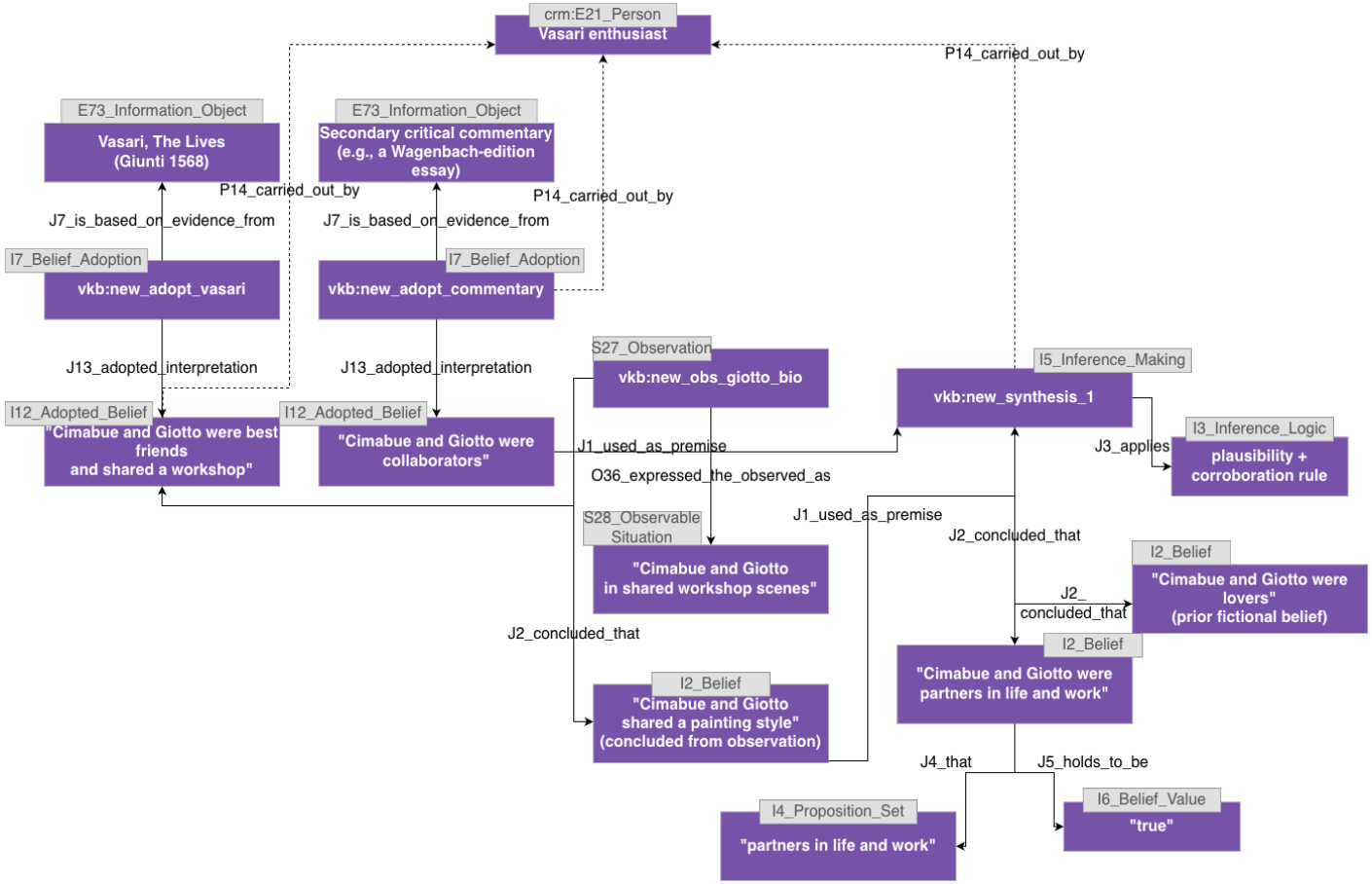


Figure 9.6: A scholarly-fidelity CRMinf encoding of the Cimabue and Giotto co-occurrence claim (Life of Cimabue), inspired by and combining elements from the canonical CRMinf specification (Doerr et al. 2023). A fictional Vasari enthusiast adopts beliefs from two `E73 Information Objects` (Vasari's *The Lives* and a secondary critical commentary) via `I7 Belief Adoption` / `J7` / `J13`, performs an `S27 Observation` on Giotto's biography that expresses an `S28 Observable Situation` via `O36`, holds a prior `I2 Belief`, and synthesizes the four resulting premises through an `I5 Inference Making` (applying an `I3 Inference Logic`) into a new `I2 Belief` whose `I4 Proposition Set` is a CRM-formalized claim about a real-world event in late-thirteenth-century Florence. The entire argumentation framework is anchored in human-actor reasoning about real-world referents; there is no class for annotating text. Compare with the Viewsari encoding of the same passage in Figure 9.7.

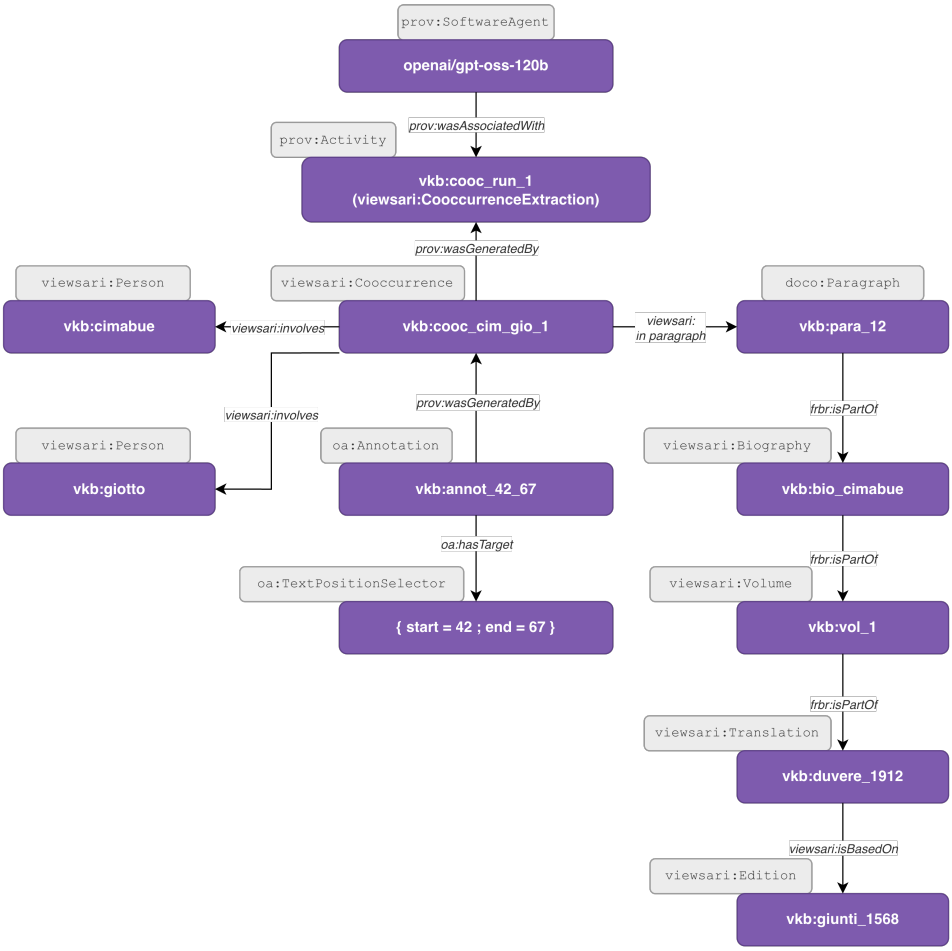


Figure 9.7: Modeling the same passage in Viewsari. The interpretive act is cast as a `prov:SoftwareAgent` executing a typed `prov:Activity` that emits a `viewsari:Cooccurrence` annotation anchored via `oa:hasTarget` to the character-offset span in paragraph 12 of the de Vere translation of the 1568 Giuntina edition. The FRBR-aligned bibliographic chain (paragraph → biography → volume → translation → edition) preserves the exact editorial location of the textual mention. Compare with Figure 9.6. The LLM is neither a human actor nor a reasoner over referents, and the two cast choices are not interchangeable.

Every load-bearing class in the Bologna encoding is the wrong class for an ObliquER-produced annotation, and the scope statements indicate that this is the intended state of affairs rather than an oversight. The descriptive observation, then, is symmetric. The CRMinf encoding does not say what the Viewsari encoding says (that the text contains a typed mention), because that is not its subject matter; the Viewsari encoding does not say what the CRMinf encoding says (that the artists met in real life), because that is also not its subject matter. The interpretive consequence of this symmetry – that the two encodings address different ontological levels and co-reside rather than compete in a humanities KG stack – is discussed in Section 11.5.

9.3 Qualitative features: epistemic adequacy

This section instantiates the *usability-profiling* dimension presented by Gangemi et al. (2006), re-scoped from interface-pragmatic annotations (recognition, efficiency, interfacing) to the epistemic bindings that the humanities setting elevates to first-class evaluation criteria. Epistemic adequacy is the qualitative dimension of ontology evaluation that asks whether the formal representation preserves the epistemic character of the source material. For a humanities KG constructed from interpretive historical sources using generative extraction, this requires assessing whether the knowledge representation qualifies the manner of reference at the data level, records the interpretive act, preserves the provenance of prompts, and distinguishes long-tail entity status from linking failure. The four properties are identified by the theoretical framing of Chapter 2 as load-bearing for epistemic adequacy and are evaluated here as observable features of the ontology rather than as theoretical desiderata. The present section verifies the bindings categorically; their quantitative reflexes – provenance coverage, the OOKB linking distribution, and the verbosity cost – are reported in Section 9.5.4 and Section 9.5, and are not re-tabulated here.

9.3.1 Evaluation method for epistemic adequacy

The evaluation method is descriptive verification at the data level rather than a numerical metric. Each of the four properties is operationalized as a data-property binding declared by the ontology and inspected in the closed populated KG. A property satisfies the criterion if (i) the ontology declares the binding as a typed property on the appropriate class, and (ii) the populated graph realizes the binding on every relevant instance rather than on a sampled subset. The criterion is binary by design – the schema and the population either honor the binding or they do not – and is intended to complement, not replace, the quantitative provenance-coverage figures of Section 9.5.4 and the comparative modeling study of Section 9.2. The four bindings inspected in

the next subsection map, respectively, onto epistemic qualification of reference, the interpretive act, the provenance of prompts, and long-tail entity status.

9.3.2 Observable bindings in the populated graph

In the following, we elaborate on the different bindings that make the Viewsari ontology epistemically adequate, conforming to the epistemic character of the source material in preserving this layer in its formal representation.

The reference as epistemic quality. One of the bindings is given implicitly through mention typology, which is the epistemic qualification of reference.

Every mention is typed following the mention taxonomy, one of (`viewsari:ExplicitArtworkMention`, `viewsari:ImplicitArtworkMention`, `viewsari:Coreferent`, or `viewsari:GenericMention`), and the type records the *manner* of identification rather than treating all references as flat assertions. Typing a span as `ImplicitArtworkMention` encodes at the data level that the referent is recovered by contextual inference rather than by canonical naming; typing a span as `Coreferent` encodes that the identification is anaphoric; typing a span as `GenericMention` encodes that no specific referent is asserted at all. The statements of Vasari's *Lives* are therefore not collapsed into bare subject–predicate–object triples without epistemic qualification. The attributive character of the source is preserved by the typing itself, queryable on every annotation, and records *what kind* of identification was made rather than only that something was identified. As Tzitzikas et al. (2022) independently record in their survey of machine-learning approaches to CIDOC CRM, IE from text remains among the most costly and least automated tasks across the surveyed systems, and none of those systems capture the epistemic provenance of the extraction itself – a descriptive backdrop against which the present binding can be read.

The interpretive act. Every extraction is recorded as an activity (of a specific type, e.g., an NER or EL) and attributed to a documented agent (such as the LLM). The act of the agent identifying, for example, an implicit artwork mention – contextual inference rather than lexical recognition (Section 2.2) – is in this way represented transparently through a detailed provenance chain which surfaces involved components in the process, avoiding a flattening of this process into the annotation it produced.

Provenance of prompts. Each activity node carries a `viewsari:promptTemplateHash` that links to the versioned template in the project repository, together with model and agent identifiers. The probabilistic nature of LLM-mediated extraction is recorded structurally rather than ignored, and run-level variability is queryable in the data.

Long-tail entity status. Entities absent from Wikidata (Vrandečić and Krötzsch 2014) are assigned stable Viewsari-internal IRIs. They do not obtain a `owl:sameAs` resource with external linkage. Absence-from-Wikidata is thereby distinguishable at the data level from linking failure, and OOKB entities remain first-class graph nodes available for future reconciliation as knowledge-base coverage improves.

9.3.3 Descriptive grounding in the populated graph

The descriptive grounding for these bindings is the populated KG itself rather than a separate evaluation experiment. The 100 % provenance coverage result (Section 9.5.4) verifies that the activity and prompt-provenance bindings are realized on every annotation in the graph; the 12,529 OOKB artwork entities documented in the M09 evaluative reading of Table 9.21 verify that the long-tail-status binding distinguishes absence-from-Wikidata at population scale; and the mention-class typing is populated on every typed mention reported in Section 8.3. The headline counts of the populated graph live in Section 8.3; the present section draws on them as inputs to the verification rather than re-reporting them.

The interpretive reading of these components of the dissertation's central claim about modeling and making visible the construction process, and how they together close the categorial gap identified in the comparative modeling study (Section 9.2) – is taken up in Section 11.1.

9.4 ObliquER: baseline comparison and ablation

The following section defines the metrics and baselines and presents the evaluation of ObliquER on entity recognition and entity linking. Coreference resolution is omitted to focus on the primary tasks of entity recognition and entity linking; coreference resolution is reserved for future work.

9.4.1 Metrics and statistical methodology

We evaluate recognition results under the SemEval 2013-9.1 span-level metric suite introduced by Segura-Bedmar et al. (2013): *Strict* requires exact span and exact type match; *Type* gives credit to overlapping spans with correct type; and *Partial* gives half credit to overlapping spans regardless of type. We use the three-criterion suite rather than a single aggregate F1 because it decomposes recognition error into independent type and boundary contributions. A system that finds the correct span but misclassifies its type produces a different error than a system that classifies the type correctly but misbounds the span. The long-tail descriptive-paraphrase mention regime of *The Lives* produces both error modes at non-negligible rates (Section 9.4.3), and the four-way decomposition isolates them. Two matching regimes appear in the tables that follow. Coreferent and generic mentions are excluded from the evaluation (consistent with the GT design of Section 7.4.4) but are reported separately when relevant. Table 9.6 additionally reports *surface-form* matching, in which a predicted mention is scored by string match against the gold surface form (its *Strict* and *Partial* columns). The reason is a comparison with state-of-the-art zero-shot entity recognition methods, which, instead of identifying mention spans, generate mention surface forms directly. The per-criterion decomposition tables – Table 9.7, Table 9.16, Table 9.17, Table 9.18, Table 9.19, and Table 9.10 – instead report *span-level bipartite* matching across the *Strict* (exact span, exact type), *Partial* (character-overlap credit regardless of type), and *Type* (character-overlap credit with matching type) criteria.

Inter-annotator agreement and the headline criterion. A second annotator independently re-annotated a stratified sample of 25 paragraphs drawn from five biographies, both to validate annotation consistency and to calibrate which criterion best reflects recognition capability. Span agreement on mentions is strong under partial matching (F1 $= 0.83$; token-level $\kappa = 0.70$, substantial agreement on the Landis and Koch (1977) scale) but only moderate under exact-span matching (F1 $= 0.43$), and the gap reflects boundary disagreement on implicit, descriptive mentions rather than disagreement on whether a mention exists at all. Agreement on every decision that is not the span boundary is, by contrast, high. Type agreement on matched spans reaches a Cohen's κ (Cohen 1960) of 0.89 (exact) and 0.70 (partial), Wikidata-QID agreement on commonly linked spans 0.84, and OOKB-classification agreement F1 $= 0.90$. Two trained readers therefore agree on which mention exists, what type it carries, and which entity it denotes, and diverge mainly on where each span begins and ends. We read the recognition results accordingly. The Partial and Type criteria, which credit overlapping spans, are the primary measure of recognition capability, because exact-span scoring penalizes the one judgment on which even human annotators disagree; the Strict (exact-span-and-type) criterion is retained chiefly for the baseline comparison of Table 9.6, where the surface form-only baselines fail at

the categorical level of emitting an implicit span at all, so the exact-match gap there registers a qualitative capability difference rather than boundary precision.

Entity-linking metrics. We evaluate linking results at the mention level. A linked mention is a true positive if the predicted Wikidata QID matches the GT QID, or if both are flagged as OOKB. We report precision, recall, and F1 separately for KB-linked and OOKB mentions, as well as the candidate recall of the retrieval stage (Dong et al. 2023). False positives and false negatives are defined differently for each metric setting. In the KB-linked setting, any KB-linked mention identified as OOKB is counted as a false negative, and a mislink is counted as a false positive. In the OOKB setting, any OOKB mention linked to an entity is counted as a false negative, while any KB-linked mention detected as OOKB is counted as a false positive. A mislink of a KB-linked mention does not affect the OOKB metrics. The KB-linked / OOKB split is load-bearing. The Viewsari corpus's high OOKB rate means that a single aggregate score would conflate retrieval failure with long-tail discovery and would reward a linker that defaults to OOKB on any candidate-set miss. We separate the candidate-recall column from the KB-linked F1 column so that the retrieval contribution and the candidate-ranking contribution can be diagnosed independently.

Evaluation corpora. We use two corpora as the evaluation surface. The Viewsari GT corpus (232 paragraphs, 1,010 explicit-and-implicit mention instances, 848 globally-deduplicated artwork entities with an OOKB ratio of 63.4 %) is described in Section 7.4.4 (construction) and Section 8.2.2 (corpus statistics).[4] The ArtPedia transferability corpus (7,618 paragraphs, 21,774 mentions, 2,679 entities, OOKB ratio 0.26 %) is described in Section 8.2.3. The two corpora differ approximately an order of magnitude in scale and invert the OOKB skew, reporting on both isolates implicit-recognition stress at scale (ArtPedia) from long-tail-linking stress (Viewsari).

Statistical methodology. The numerical values reported in the following subsections extend the evaluation on the unified-exclusion test partition. Per-paragraph bootstrap 95 % confidence intervals (paragraph-level resampling with $N = 10{,}000$ replicates, following Dror et al. 2018) are reported in Table 9.5 for the *Strict* F1 under surface-form matching and the *Strict* F1 under span-level matching of every recognition configuration on the Viewsari GT; corresponding intervals for

[4] The corresponding per-biography-sum counts – entities recurring across k biographies contributing k to the total, as in the accompanying conference paper – differ from the globally-deduplicated figures used here; the globally-deduplicated convention is preferred throughout this chapter for direct comparability with the populated-KG figures of Section 9.5.

the Type and Partial criteria are an order of magnitude tighter and are not separately tabulated. The resampling unit is the paragraph because the per-paragraph mention count is the natural denominator on a sample of 232 Viewsari paragraphs and because mention-level resampling would break the within-paragraph dependence between coreferent and antecedent mentions.

Table 9.5: Bootstrap 95 % confidence intervals on the strictest-criterion F1 for every recognition configuration reported on the Viewsari GT in this chapter. The *Strict* criterion (exact span and type) under surface-form matching and the *Strict* criterion (exact span and type) under span-level bipartite matching, computed via per-paragraph resampling with $N = 10{,}000$ replicates following Dror et al. (2018).

		Strict (surface-form)		**Strict (span-level)**	
Configuration	**Model**	**F1**	**95 % CI**	**F1**	**95 % CI**
A1: Monolith prompt	`gpt-oss-120b`	0.279	[0.250, 0.308]	0.284	[0.255, 0.314]
A1: Monolith prompt	`gemma-4-31b-it`	0.305	[0.270, 0.336]	0.300	[0.266, 0.332]
A2: No domain-specific content	`gpt-oss-120b`	0.239	[0.206, 0.273]	0.249	[0.216, 0.283]
A2: No domain-specific content	`gemma-4-31b-it`	0.227	[0.194, 0.261]	0.224	[0.190, 0.257]
A3: Zero-shot ontology (k=0)	`gpt-oss-120b`	0.330	[0.299, 0.360]	0.330	[0.299, 0.361]
A3: Zero-shot ontology (k=0)	`gemma-4-31b-it`	0.326	[0.287, 0.362]	0.323	[0.284, 0.359]
A4: Few-shot, k=1	`gpt-oss-120b`	0.304	[0.271, 0.337]	0.307	[0.274, 0.339]
A4: Few-shot, k=1	`gemma-4-31b-it`	0.340	[0.305, 0.376]	0.336	[0.301, 0.371]
A4: Few-shot, k=3	`gpt-oss-120b`	0.316	[0.285, 0.348]	0.316	[0.285, 0.346]
A4: Few-shot, k=3	`gemma-4-31b-it`	0.361	[0.330, 0.394]	0.358	[0.327, 0.391]
A4: Few-shot, k=5	`gpt-oss-120b`	0.342	[0.313, 0.373]	0.344	[0.313, 0.374]
A4: Few-shot, k=5	`gemma-4-31b-it`	0.374	[0.337, 0.411]	0.371	[0.336, 0.409]
Few-shot prompt variant	`gpt-oss-120b`	0.302	[0.266, 0.336]	0.310	[0.274, 0.345]
CoT prompt	`gpt-oss-120b`	0.322	[0.286, 0.356]	0.327	[0.292, 0.361]
Ontology-guided, k=2 (prod.)	`gpt-oss-120b`	**0.334**	[0.302, 0.367]	**0.334**	[0.300, 0.368]
Ontology-guided, k=2 (prod.)	`gemma-4-31b-it`	**0.364**	[0.326, 0.402]	**0.361**	[0.323, 0.399]

The remaining subsections report results. Section 9.4.2 defines the baselines; Section 9.4.3 reports the entity-recognition comparison; Section 9.4.4 the EL comparison; Section 9.4.5 the cross-model evaluation; Section 9.4.6 the ablation study; and Section 9.4.7 the ArtPedia transferability assessment. The interpretive reading of these results against the central claim and against RQ2 is taken up in Section 11.3.

9.4.2 Baseline definitions for recognition and linking

We compare ObliquER against several baselines, two for entity recognition and three for entity linking, each explained in the following paragraphs.

B1 — UniversalNER. UniversalNER 7B (UniNER-7B) (Zhou et al. 2024c) is the instruction-tuned distilled NER model used by Santini et al. (2023) for artwork and motif detection on the comparable ExtrART corpus;[5] we choose B1 for direct comparability with the prior art on Vasari rather than for its absolute performance, and we do not re-quote the specific F1 figures reported in that work on ExtrART, since the comparison run is against a different corpus and the prior numbers are not directly transferable. We rerun UniNER-7B on the Viewsari GT using the same metrics used in the UniNER publication (strict and partial surface-form matching) in Section 9.4.1. The implicit-mention recall is structurally zero, regardless of how the model performs on the explicit subset.

B2 — GLiNER. GLiNER2 (Zaratiana et al. 2025) extends GLiNER (Zaratiana et al. 2024) to a unified schema-based extraction framework that supports NER, text classification, structured data extraction, and relation extraction within a single 205M-parameter encoder model (a 340M-parameter variant is also available). GLiNER2 retains the zero-shot label-conditioning architecture of its predecessor and additionally admits entity descriptions – natural-language definitions associated with each label – which are passed alongside the label string at inference time. B2 is the strongest encoder-based zero-shot NER baseline that admits descriptions, and therefore the strongest control for separating the contribution of the description from the contribution of the generative decoder in ObliquER. Two GLiNER2 configurations are run on the Viewsari GT:

- **B2-label**. The label `"artwork"` is passed without a description, mirroring the input regime of the earlier GLiNER baseline. This configuration isolates the model update's contribution from the configuration change.
- **B2-desc**. The label `"artwork"` is passed together with the natural-language description *"A creative or artistic product, including paintings, sculptures, architectural constructions, altarpieces, or decorative artworks. Refers to a specific artwork, not a category or a process describing a person's interest or talent."*, which corresponds verbatim to the entity-definition block of the full ObliquER prompt (Section 7.4.8.3).

[5] Corresponding GitHub repository: `https://github.com/ISE-FIZKarlsruhe/vasari_nlp/tree/main`.

Under B2-label, GLiNER2 receives the same label-only input as the older GLiNER baseline; under B2-desc, it receives the same conceptual specification that drives ObliquER's recognition stage, but without the pipeline staging, few-shot examples, or generative decoding. Like B1, GLiNER2 operates on surface form-level, so implicit-mention recall is structurally zero. The pairing of B2-label and B2-desc is therefore neither a hyperparameter sweep nor an ablation in the ObliquER sense (Section 9.4.6); it is a controlled comparison in which the two configurations differ in exactly one input variable.

B3 — ReFinED (EL). ReFinED (Ayoola et al. 2022) is an end-to-end zero-shot EL model trained to resolve mentions against the full Wikidata entity set (>30M entities). The choice of B3 is motivated by ReFinED's status as the current state of the art for zero-shot Wikidata linking under a span-based candidate-generation regime, against which ObliquER's artist-constrained retrieve-then-rank strategy is to be compared. ReFinED does not handle OOKB detection of implicit mentions as positive outputs of its decision procedure; its OOKB column is populated by fallback whenever the candidate set is empty. This architectural property is named here because it bears on how the OOKB column of the linking-comparison table is to be read (Section 9.4.4), not as a result claim about ReFinED's performance.

B4 — BLINK (EL). BLINK (Wu et al. 2020) is the canonical two-stage dense-retrieval EL system. In BLINK, a bi-encoder independently embeds the mention-in-context and the entity description, retrieves a candidate set in a shared dense space, and a cross-encoder concatenates the mention and entity text and re-ranks the candidates. This approach is structurally similar to ObliquER's retrieve-then-rank design. BLINK performs in-KB evaluation only and defers NIL prediction to future work; it therefore carries no OOKB category in its decision space, and the OOKB column is zero, not a structural absence rather than a linking failure (Section 9.4.4).

B5 — GENRE (EL). GENRE (De Cao et al. 2021) represents the generative paradigm for entity retrieval. Instead of scoring a candidate set, it generates the target entity's canonical name token by token in an autoregressive pass, decoding under a constrained beam search that restricts each step to a prefix trie of valid identifiers. Because GENRE is end-to-end generative and consults no candidate set, candidate recall is not defined for it on Viewsari (reported as "–" in Table 9.11); and because its constrained decoder can emit only names present in the KB, it likewise produces no OOKB output, so its OOKB column is zero.

9.4.3 Comparison results: entity recognition

The headline P/R/F1 figures under *Strict* and *Partial* matching for ObliquER and the two zero-shot baselines (UniNER-7B, GLiNER2) on both Viewsari and ArtPedia are reported in Table 9.6; the standalone ObliquER recognition behavior decomposed across all three SemEval evaluation criteria (Strict / Type / Partial) is reported in Table 9.7.[6]

Table 9.6 reports an ObliquER *Strict* F1 of 0.334 on Viewsari (All) against 0.09 for GLiNER2 and 0.02 for UniNER-7B. On the Implicit subset, the surface form-only baselines return F1 $= 0.02$ and F1 $= 0.00$ respectively, while ObliquER returns F1 $= 0.321$. The per-criterion table for ObliquER (Table 9.7) separates the two error sources directly. Holding type aside, relaxing the span match from exact to overlapping lifts F1 from *Strict* 0.334 to *Partial* 0.697 on Viewsari ($0.321 \rightarrow 0.608$ on ArtPedia), so boundary detection rather than entity discovery accounts for the bulk of the exact-match shortfall; adding the type requirement back at overlapping spans moves *Partial* 0.697 to *Type* 0.642 ($0.608 \rightarrow 0.587$ on ArtPedia), a small drop that bounds type-classification error. Boundary error therefore dominates type error, consistent with the inter-annotator agreement of Section 9.4.1. To examine where the recognition errors fall *within* the prescribed Viewsari mention typology – a level of detail not available for the surface form-only baselines – Table 9.8 reproduces the confusion matrix of the production configuration on the updated Viewsari GT. Rows are gold mention types; columns are the predicted types, with an additional *Missed* column for gold mentions for which the model produced no overlapping span.

The Implicit row of Table 9.8 reports the 771 implicit gold mentions on Viewsari as 569 correctly typed *Implicit*, 57 typed as *Explicit* on the correct span, 12 typed as *Coreferent*, 6 typed as *Generic*, and 127 missed (≈16 %). The Explicit row reports 209 explicit gold mentions: 158 correctly typed *Explicit*, 14 typed *Implicit*, 5 typed *Coreferent*, and 32 missed. These 771 implicit and 209 explicit gold mentions (980 in total) are the recognition evaluation's *unified-exclusion partition*. The 1,010 implicit-and-explicit mentions of the corrected GT corpus (Section 7.4.7)

[6] All Viewsari ObliquER numbers in the tables and prose of this chapter are evaluated on the *unified-exclusion test partition* that removes the few-shot demonstration and modeling-example paragraphs (Ghirlandaio 251, Cimabue 1, Giotto 112, Titian 250, Brunelleschi 206, Lippi 9) from the gold GT to avoid prompt/test leakage across the recognition prompt variants and the A4 few-shot-count ablation; the partition rationale is described in full in the preamble of Section 9.4.6. The numbers were computed on the full annotation including these paragraphs; the resulting figures differ by margins on the order of ± 0.01–0.03 per cell, and the qualitative findings of the chapter – the implicit-mention recognition gap over surface form-only baselines, the cross-model invariance of the prompt-architecture hypothesis, the rank-ordering of the architectural pillars, and the non-monotone few-shot curve – are preserved across both partitions. The UniNER-7B and GLiNER2 baseline rows of Table 9.6 have been re-run on the unified-exclusion partition under the updated GT. The ArtPedia ObliquER row of Table 9.6 (‡) is evaluated on the analogous ArtPedia partition that excludes the 4 mentions contained in the paragraph used as a few-shot demonstration for the ArtPedia recognition prompt (Section 7.4.8.3); the resulting evaluation count is 21,770 mentions out of the 21,774 in the full ArtPedia annotation (Section 8.2.3), a 0.018 % exclusion that does not perceptibly affect the headline figures.

Table 9.6: Strict and partial surface-form matching for ObliquER versus zero-shot NER baselines on the Viewsari GT and on ArtPedia. *Strict* = exact span and type; *Partial* = half credit for overlapping spans regardless of type. Viewsari rows are broken down by mention subset (Explicit, Implicit, All); ArtPedia is reported overall, since its annotation schema collapses to two mention types. The Viewsari ObliquER rows report the production ontology-guided 2-shot configuration under `openai/gpt-oss-120b`; the UniNER-7B and GLiNER2 baseline rows are re-run on the same unified-exclusion partition as the ObliquER rows. See the partition footnote on the lead paragraph of this subsection for the evaluation-set definition shared across all Viewsari cells of this chapter.

		Strict			Partial		
Dataset	**Method (subset)**	**P**	**R**	**F1**	**P**	**R**	**F1**
Viewsari	B1 UniNER-7B (Expl.)	0.03	0.07	0.04	0.05	0.14	0.08
	B1 UniNER-7B (Impl.)	0.00	0.00	0.00	0.19	0.08	0.11
	B1 UniNER-7B (All)	0.02	0.02	0.02	0.10	0.09	0.09
	B2 GLiNER2 (Expl.)	0.29	0.20	0.23	0.35	0.24	0.29
	B2 GLiNER2 (Impl.)	0.13	0.01	0.02	0.44	0.04	0.08
	B2 GLiNER2 (All)	0.23	0.05	0.09	0.38	0.09	0.14
	ObliquER (Expl.)	**0.303**	**0.493**	**0.375**	**0.396**	**0.643**	**0.490**
	ObliquER (Impl.)	**0.285**	**0.367**	**0.321**	**0.440**	**0.567**	**0.495**
	ObliquER (All)	**0.289**	**0.394**	**0.334**	**0.428**	**0.583**	**0.494**
ArtPedia	B1 UniNER-7B	0.12	0.07	0.09	0.22	0.12	0.16
	B2 GLiNER2	0.19	0.06	0.09	0.29	0.09	0.14
	ObliquER[‡]	**0.330**	**0.308**	**0.319**	**0.489**	**0.457**	**0.472**

minus the 30 that fall in the six held-out demonstration paragraphs excluded to prevent prompt/test leakage (Section 9.4.6). The same 980-mention partition underlies every span-level recognition table in this section. The off-diagonal cells record a type-misclassification asymmetry of 57 implicit → predicted *Explicit* against 14 explicit → predicted *Implicit*. The surface form-only baselines UniNER-7B and GLiNER2 return zero on the implicit column of Table 9.6, since they emit no implicit-typed predictions; on the explicit subset, ObliquER's *Strict* F1 of 0.375 exceeds GLiNER2's 0.23 and UniNER-7B's 0.04.

The ArtPedia confusion matrix (Table 9.9) reports the full Viewsari prediction typology plus a missed column. Reading the matrix in the same direction as for Viewsari (rows = gold, columns = predicted), the 19,549 implicit gold mentions are split into 10,533 correctly typed *Implicit*, 415 typed as *Explicit*, 29 typed as *Generic*, 0 typed as *Coreferent*, and 8,572 missed; the 2,221 explicit gold mentions are split into 1,907 correctly typed *Explicit*, 27 typed as *Implicit*, 1 typed as *Generic*, 0 typed as *Coreferent*, and 286 missed. The off-diagonal cells record 415 implicit → predicted *Explicit* against 27 explicit → predicted *Implicit*, so the type-misclassification

Table 9.7: ObliquER entity-recognition results under *span-level bipartite matching*, decomposed across the exact, partial, and type criteria. *Strict* requires an exact span and exact type; *Partial* gives character-overlap credit regardless of type; *Type* gives character-overlap credit with matching type. The Viewsari columns report the production ontology-guided 2-shot configuration under `openai/gpt-oss-120b`; the ArtPedia columns report the few-shot configuration.

	Viewsari			ArtPedia		
Eval. Type	**P**	**R**	**F1**	**P**	**R**	**F1**
Strict	0.292	0.391	0.334	0.330	0.312	0.321
Partial	0.609	0.815	0.697	0.625	0.592	0.608
Type	0.561	0.751	0.642	0.603	0.571	0.587

Table 9.8: Per-mention-type partial recognition behavior of ObliquER on the Viewsari GT. Rows: gold types; columns: prediction. Generic and Coreferent gold rows are excluded because they were not evaluated (out of scope). Computed on the updated Viewsari GT under the unified-exclusion partition (see the footnote of Section 9.4.3).

Gold ↓ \ Pred. →	**Gen.**	**Impl.**	**Coref.**	**Expl.**	**Missed**
Implicit	6	569	12	57	127
Explicit	0	14	5	158	32

direction is the same as on Viewsari, with the same asymmetry magnified by ArtPedia's order-of-magnitude larger implicit subset.

Entity recognition error analysis

The two confusion matrices and the per-criterion table together expose three dominant recognition failure modes, reported here as a descriptive taxonomy of where the matrices' off-diagonal mass sits.

- **Highly descriptive implicit mentions confused with general artistic descriptions.** False positives in which a passage about an artist's practice (*"his skill in fresco"*, *"the manner of painting employed in Florence"*) is predicted as an implicit artwork mention. This is the most frequent false-positive class on Viewsari and accounts for the residual precision gap on the implicit subset.
- **Boundary detection for implicit mentions.** Implicit gold spans on Viewsari are typically multi-clause descriptive phrases (*"a panel for the friars of Santa Croce representing the Last Supper"*); the model frequently identifies the correct entity type but selects a sub-span or super-span. The failure mode is invisible to *Type*-only matching but suppresses *Strict* F1

Table 9.9: Per-mention-type partial recognition behavior of the few-shot configuration on the ArtPedia evaluation set. Rows: gold; columns: prediction. The ArtPedia annotation schema collapses gold-side to two mention types (Implicit, Explicit), with no Generic or Coreferent gold rows; the prediction side carries the full Viewsari four-way typology plus a *Missed* column for gold mentions for which the model produced no overlapping span.

Gold ↓ \ Pred. →	Gen.	Impl.	Coref.	Expl.	Missed
Implicit	29	10533	0	415	8572
Explicit	1	27	0	1907	286

substantially, and accounts for the bulk of the *Type* F1 $= 0.642$ vs. *Strict* F1 $= 0.334$ gap on Viewsari (Table 9.7).[7]

- **Entity-type misclassification.** The 57 implicit → *Explicit* and 14 explicit → *Implicit* cases on Viewsari (Table 9.8) and the 415 implicit → *Explicit* and 27 explicit → *Implicit* cases on ArtPedia (Table 9.9) record the same descriptive-paraphrase / canonical-title boundary at the typological level. The ablation in Section 9.4.6 reports how the aggregate Strict-, Type-, and Partial-F1 figures shift under prompt-design variations; per-mention-type confusion matrices for each ablation condition are not reported.

The interpretive reading of these three error modes – the causal account of the ObliquER-vs-baseline gap, the relation between the descriptive-paraphrase boundary and the corpus-level annotation conventions, and the candidate remediations across prompt, few-shot, and annotation-guideline dimensions – is taken up in Section 11.3.

Recognition variance across the 16 biographies

The headline figures of Table 9.6 and Table 9.7 aggregate over the entire Viewsari GT. The per-biography breakdown of the production ontology-guided 2-shot configuration under `openai/gpt-oss-120b` (Table 9.10) shows that the aggregate hides substantial variance. Across the 16 biographies, Type F1 spans a range of 0.331, from 0.768 on Tiziano da Cadore

[7] The boundary-detection failure mode is also visible when the recall denominator is split by mention-depth scope. The Viewsari GT distinguishes 847 outer (top-level) mentions from 133 inner (nested) mentions, where an inner mention is a mention whose span sits inside the span of another mention – a canonical title embedded in a longer descriptive paraphrase, for example. Under the production configuration, exact-span recall is *higher* on the inner subset ($R_{\text{str}} = 0.511$) than on the outer subset ($R_{\text{str}} = 0.372$), the counter-intuitive direction relative to standard "inner mentions are harder" folklore. Type-conditional recall ($R_{\text{typ}} = 0.782$ inner vs 0.763 outer) and partial-overlap recall ($R_{\text{par}} = 0.880$ inner vs 0.849 outer) are comparable across the two subsets. The asymmetry reflects the composition of nested cases on Viewsari. Inner mentions are predominantly short canonical names embedded in longer descriptive outer spans, so exact-span match is easier on the inner side and harder on the outer side, where the model must reproduce a multi-clause descriptive boundary.

down to 0.437 on Michelangelo Buonarroti. The biographies cluster in two visible regimes – fifteen biographies between 0.589 (Brunelleschi) and 0.768 (Tiziano), with Michelangelo isolated at 0.437, more than 0.15 below the next-lowest biography (Brunelleschi) and 0.20 below the chapter-wide Type F1 of 0.642.

Table 9.10: Per-biography recognition behavior of the production ontology-guided 2-shot configuration on the Viewsari GT under `openai/gpt-oss-120b` and span-level bipartite matching, ranked by Type F1 (descending). The n column gives the gold-mention count per biography. The bottom row reproduces the chapter-wide aggregate.

Biography	n	**Strict F1**	**Type F1**	**Partial F1**
Tiziano da Cadore	66	0.304	**0.768**	0.783
Antonio and Piero Pollaiuolo	64	0.422	0.766	0.844
Leon Batista Alberti	35	0.474	0.737	0.737
Sandro Botticelli	57	0.439	0.732	0.764
Jacopo da Pontormo	40	0.269	0.731	0.750
Andrea Verrocchio	73	0.453	0.721	0.779
Alesso Baldovinetti	22	0.264	0.717	0.792
Gherardo	22	0.238	0.714	0.762
Pietro Cavallini	40	0.376	0.706	0.753
Giovanni Cimabue	58	0.310	0.676	0.704
Cosimo Rosselli	33	0.333	0.667	0.718
Domenico Ghirlandajo	112	0.442	0.667	0.708
Giotto	89	0.335	0.609	0.680
Filippo Lippi (Filippino)	65	0.291	0.596	0.755
Filippo Brunelleschi	76	0.325	0.589	0.619
Michelagnolo Buonarroti	128	0.186	*0.437*	0.508
All biographies (chapter-wide)	**980**	**0.334**	**0.642**	**0.697**

The Michelangelo outlier carries methodological significance beyond its single-biography F1 deficit. The GT construction described in Section 7.4.4 (specifically Section 7.4.4.2) treats Michelangelo's biography as the anchor case for prompt and schema development, retaining 25 annotated paragraphs against the 10–20 cap applied to the remaining biographies; the schema was iteratively tuned against this biography during the development pass. The production prompt is therefore the configuration on which Michelangelo's biography has had the most direct influence at design time, and it is also the biography on which the prompt performs the worst at evaluation time. The variance is not solely a recall problem. Michelangelo's predicted-mention count ($n_{\text{pred}} = 238$) substantially exceeds the gold count ($n = 128$), so the dominant error mode is overprediction. The model emits roughly 1.86 mentions per gold mention in this biography, compared with the

chapter-wide ratio of 1.34 predicted-to-gold (1,312 predicted against 980 gold). The interpretive reading of this within-corpus variance – the relation between Michelangelo's biographical density of descriptive iconographic prose, the prompt's saturation against that prose at design time, and the resulting paragraph-level over-prediction at evaluation time – is taken up in Section 11.3.

9.4.4 Comparison results: entity linking

Table 9.11 compares ObliquER's EL performance against the linkers ReFinED (B3), BLINK (B4), and GENRE (B5) on the Viewsari GT. P/R/F1 are reported separately for KB-linked mentions (gold mentions whose target Wikidata QID is present in the candidate set) and OOKB (OOKB) mentions (gold mentions for which no canonical Wikidata entry exists). Candidate recall is the fraction of gold KB-linked mentions for which the correct identifier appears among the candidates supplied to the linker; GENRE is generative end-to-end and does not use a candidate set, so candidate recall is not applicable.

Table 9.11: EL on the Viewsari GT under gold-mention inputs, broken down across all, explicit, and implicit mention subsets. P/R/F1 are reported separately for KB-linked and OOKB mentions; CR denotes candidate recall. GENRE is generative end-to-end and does not use a candidate set, so CR is reported as "–".

		KB-linked			OOKB			
Model	**Subset**	**P**	**R**	**F1**	**P**	**R**	**F1**	**CR**
ObliquER	All	**0.417**	**0.446**	**0.431**	0.740	**0.703**	**0.721**	**0.671**
	Explicit	**0.640**	**0.435**	**0.518**	0.271	0.703	**0.391**	**0.696**
	Implicit	**0.333**	**0.455**	**0.385**	0.834	**0.703**	0.763	**0.652**
ReFinED	All	0.284	0.092	0.139	0.621	0.928	0.744	0.110
	Explicit	0.430	0.181	0.255	0.183	0.788	0.297	0.207
	Implicit	0.097	0.024	0.039	0.704	0.936	**0.804**	0.036
BLINK	All	0.094	0.223	0.132	0.000	0.000	0.000	0.425
	Explicit	0.299	0.351	0.323	0.000	0.000	0.000	0.654
	Implicit	0.038	0.126	0.059	0.000	0.000	0.000	0.251
GENRE	All	0.055	0.131	0.078	0.000	0.000	0.000	–
	Explicit	0.190	0.223	0.205	0.000	0.000	0.000	–
	Implicit	0.019	0.061	0.028	0.000	0.000	0.000	–

Table 9.11 reports an ObliquER KB-linked F1 of 0.431 on Viewsari (All), roughly three times the best baseline (0.139 ReFinED, 0.132 BLINK, 0.078 GENRE), and ObliquER leads on both

the Explicit and Implicit subsets. ObliquER's artist-constrained retrieval strategy (Section 7.4.3) operates over a pool of artworks attributed to the biography's protagonist via Wikidata's `wdt:P170` creator property rather than over surface-form matches against the Wikidata index, which is the structural reason its candidate recall (0.652 implicit, 0.696 explicit) exceeds the surface-form baselines by an order of magnitude.

In the OOKB column, ObliquER and ReFinED report comparable F1 on the All and Implicit subsets (0.721 vs. 0.744 on All, 0.763 vs. 0.804 on Implicit – both within $\tilde{0}.04$ of each other); on the Explicit subset, where mis-routing to OOKB indicates a linking failure rather than a long-tail entity, ObliquER reports 0.391 against ReFinED's 0.297. The two systems route to OOKB through different mechanisms. ReFinED's OOKB classification results from an empty candidate set returned by its surface-form dictionary, whereas ObliquER's OOKB classification is the output of the candidate-ranking LLM operating on a non-empty candidate set (Section 7.4.3). The Explicit-subset gap is the empirical signature that the discriminative OOKB mechanism does measurable work where it matters most, on canonical-title mentions for which an OOKB call is a linking failure rather than a long-tail discovery. BLINK and GENRE return zero across the OOKB column because they have no OOKB representation in their decision space.

ArtPedia entity linking

Table 9.12 reports the same EL comparison on the ArtPedia test split, evaluated against the gold Wikidata identifiers harvested via the title-to-QID mapping described in Section 7.4.10.1, broken down by mention subset (All, Explicit, Implicit) to mirror the Viewsari decomposition of Table 9.11. The OOKB slice on ArtPedia is small – the dataset is derived from Wikipedia and almost every annotated artwork resolves to an existing Wikidata QID (Section 8.2.3) – so the OOKB column is read here as an OOKB-detection sanity check rather than as a contested comparison. The artist-constrained pool used by ObliquER's retriever stack is also better populated on ArtPedia than on Viewsari, since Wikipedia-grounded artworks more frequently carry the `wdt:P170` creator property that seeds the pool. The four-mention ArtPedia partition exclusion described in the footnote on the lead paragraph of Section 9.4.3 applies to this table as well.

The headline figure is an ObliquER KB-linked F1 of 0.733 on ArtPedia (All) against 0.168 (ReFinED), 0.162 (BLINK), and 0.162 (GENRE), with candidate recall CR $= 0.965$ for ObliquER against $0.098/0.277/0.219$ for the three baselines. The 0.965 CR figure sits well above the Viewsari row of Table 9.11 (CR $= 0.671$ on All), which is the structural reason ArtPedia is a less binding test of the retrieval stage than Viewsari is. ReFinED's KB-linked precision of 0.940 paired with recall of 0.092 is the empirical signature of its empty-candidate-set fallback mechanism. When the surface-form dictionary returns a candidate set, the linker is highly accurate; when it

Table 9.12: EL on the ArtPedia test split under gold-mention inputs, broken down across All, Explicit, and Implicit mention subsets. P/R/F1 are reported separately for KB-linked and OOKB mentions; CR denotes candidate recall. GENRE is generative end-to-end and does not use a candidate set, so its CR row reports the post-decoding correctness rate rather than a retrieval-set candidate recall.

		KB-linked			OOKB			
Model	**Subset**	**P**	**R**	**F1**	**P**	**R**	**F1**	**CR**
ObliquER	All	0.775	**0.696**	**0.733**	**0.006**	0.167	**0.012**	**0.965**
	Explicit	0.920	0.902	0.911	0.000	0.000	0.000	**0.961**
	Implicit	0.757	**0.673**	**0.713**	**0.006**	0.167	**0.012**	**0.965**
ReFinED	All	**0.940**	0.092	0.168	0.004	**1.000**	0.009	0.098
	Explicit	**0.939**	0.908	0.924	0.000	0.000	0.000	**0.961**
	Implicit	**1.000**	0.002	0.004	0.004	**1.000**	0.009	0.003
BLINK	All	0.162	0.162	0.162	0.000	0.000	0.000	0.277
	Explicit	0.915	**0.915**	0.915	0.000	0.000	0.000	0.941
	Implicit	0.079	0.079	0.079	0.000	0.000	0.000	0.204
GENRE	All	0.162	0.162	0.162	0.000	0.000	0.000	0.219
	Explicit	**0.948**	**0.948**	**0.948**	0.000	0.000	0.000	0.954
	Implicit	0.075	0.075	0.075	0.000	0.000	0.000	0.138

returns nothing, the mention is routed to OOKB, which on ArtPedia is the dominant failure mode because most ArtPedia mentions are KB-resolvable but not surface-form-canonical.

The subset decomposition makes the regime distinction precise. On the Explicit subset, all four systems sit in the narrow band F1 $\in [0.911, 0.948]$, so explicit-mention EL on Wikipedia-grounded artworks is effectively saturated for the current generation of systems, and the comparison is non-discriminative at that scale. The Implicit subset, by contrast, separates the four systems by an order of magnitude. ObliquER reaches KB-linked F1 $= 0.713$ against 0.004 (ReFinED), 0.079 (BLINK), and 0.075 (GENRE), the candidate-recall column giving the structural reason (0.965 for ObliquER against $0.003/0.204/0.138$). The artist-constrained retrieval strategy (Section 7.4.3) recovers the correct candidate at near-saturation rates on implicit mentions because the bound between the biography's protagonist and the candidate pool is preserved across the corpus shift from Renaissance biographical prose to Wikipedia-grounded visual description; the surface-form-based linkers cannot bind the implicit-mention surface form to its referent, and the linker-level failure compounds the retrieval-level failure.

ObliquER's KB-linked F1 of 0.713 on the ArtPedia Implicit subset sits substantially above its 0.385 on the Viewsari Implicit subset of Table 9.11. The interpretive consequence is precise.

Under stronger surface-form anchoring (Wikipedia titles and templated visual descriptions) the implicit-recognition surface translates more cleanly into the implicit-linking surface, because the residual ambiguity once the mention is correctly typed is lower on ArtPedia than on Vasari's biographical prose, where the implicit mention typically refers to a work that is either lost, lacks a canonical title, or is identified by descriptive periphrasis only. The ObliquER advantage over the surface form-only baselines is therefore preserved across the Viewsari→ArtPedia shift and amplified in absolute level, while the relative gap to ReFinED widens (Viewsari Implicit: 0.385 vs 0.039; ArtPedia Implicit: 0.713 vs 0.004, a factor of $\sim$180 $\times$).

OOKB columns are near-zero across all systems on ArtPedia, consistent with the small absolute number of true OOKB items in this corpus (Section 8.2.3). ObliquER's OOKB F1 of 0.012 on the All and Implicit subsets is the highest among the four systems, but the absolute number is small enough that this column is best read as a sanity check on OOKB detection rather than as a contested benchmark. The architecture's first-class OOKB output does not collapse where OOKB is rare, but it also lacks the room to display the long-tail-discovery property it exhibits on Viewsari. ReFinED's OOKB recall of 1.000 is again the empty-candidate-set fallback of Table 9.11 – it predicts OOKB by default when the dictionary fails – so its OOKB F1 is dominated by the precision component (0.004).

End-to-end pipeline: recognition → linking on Viewsari

Table 9.11 and Table 9.12 report EL under *gold-mention* inputs, where each linker is asked to assign an identifier to every mention from the manually annotated GT, so the linking stage is evaluated in isolation from any recognition error. To complement this with an integrated-pipeline view, Table 9.13 reports the end-to-end behavior on Viewsari, with mentions produced by the ObliquER recognition stage rather than read from the GT and then fed directly into the same linking stack.

Table 9.13: End-to-end EL on the Viewsari GT: mentions produced by the ObliquER NER stage feed directly into the EL stage, so the reported numbers reflect both recognition and linking errors. P/R/F1 are reported separately for KB-linked and OOKB mentions; CR denotes candidate recall.

	KB-linked			OOKB			
Subset	**P**	**R**	**F1**	**P**	**R**	**F1**	**CR**
All	0.280	0.400	0.330	0.820	0.690	0.750	0.590
Explicit	0.410	0.400	0.410	0.580	0.590	0.580	0.630
Implicit	0.240	0.410	0.300	0.860	0.690	0.770	0.550

Table 9.13 reports KB-linked F1 figures of 0.410 (Explicit) and 0.300 (Implicit) under predicted mentions, against 0.518 and 0.385 under gold mentions in Table 9.11. Candidate recall shifts in the same direction, from 0.671 / 0.696 / 0.652 (All / Expl. / Impl., gold mentions) to 0.590 / 0.630 / 0.550 (predicted mentions). OOKB F1 reads 0.750 (All) and 0.770 (Implicit) under predicted mentions, against 0.721 (All) and 0.763 (Implicit) under gold mentions. The Explicit OOKB column shifts in the opposite direction to the KB-linked column, with precision rising from 0.271 to 0.580 and F1 from 0.391 to 0.580.

A symmetric ArtPedia end-to-end evaluation is not reported here. ArtPedia's single-artwork-per-paragraph annotation makes precision interpretation ambiguous under NER errors, since multiple ObliquER-detected mentions in a paragraph would be scored against a single gold mention, conflating recognition and linking errors with annotation undercounting.

Retrieval-query representation comparison: structured vs. context query

Table 9.14 reports the candidate-recall consequence of the query-rewriting stage described in Section 7.4.9.4. The two production retrievers – the bi-encoder over the prebuilt global Wikidata index ("Prebuilt WD") and the bi-encoder over the artist-constrained pool ("Author pool") – are each run twice on the ArtPedia validation set under two query encodings: *Structured*, in which the query is the LLM-generated structured metadata record (title, description, creator, depicts, location, material, etc.), and *Context*, in which the query is the raw mention surrounded by its paragraph context, encoded directly by the bi-encoder. The metric is candidate recall at the top-16 cut (CR@16), matched to the per-retriever budget retained by the greedy search of Section 7.4.9.4. The comparison is grouped here with the EL results rather than with the broader ablation study of Section 9.4.6, since the variable under comparison is intrinsic to the EL retrieval stage rather than to the ObliquER prompt architecture.

Table 9.14: Input-representation ablation. Candidate recall (CR@16) for the two dense retrievers under two query encodings, evaluated on the ArtPedia test partition. *Structured* = the LLM-generated structured metadata record used as the query; *Context* = the raw mention embedded in its surrounding paragraph context.

Retriever	Query	Overall	Explicit	Implicit
Prebuilt WD	Structured	**0.333**	**0.083**	**0.500**
	Context	0.183	0.021	0.292
Author pool	Structured	**0.308**	**0.104**	**0.444**
	Context	0.092	0.000	0.153

The structured query yields a higher CR@16 across all six cells in Table 9.14. In the Overall column, Prebuilt WD reports CR 0.333 (structured) against 0.183 (context), and Author pool reports CR 0.308 against 0.092. The largest relative gap is in the Implicit subset for the Author pool (0.444 against 0.153, a roughly threefold lift); the Explicit subset under the Author pool reaches CR $= 0.000$ with the context query, compared to 0.104 with the structured query. The structured query is the production-pipeline default (Section 7.4.9.4); the context query is reported here as the comparison baseline. The interpretive reading of the lift – why the structured query carries the implicit-subset gap, why the context query collapses to zero on the Explicit subset under the Author pool, and what these consequences imply for the central empirical claim of the dissertation – is taken up in Section 11.3.

Entity linking error analysis

A targeted error analysis of the Viewsari KB-linked predictions in Table 9.11 and Table 9.12 identifies three dominant failure modes, presented here as a descriptive taxonomy that maps the off-diagonal mass to distinct stages of the pipeline.

- **Retrieval miss masquerading as OOKB.** Mentions are classified as OOKB when the correct knowledge-base entity is missing from the candidate set returned by the retrievers. These cases are formally OOKB *predictions* but not OOKB *discriminations*. The ranker had no in-KB option to choose from. On the Viewsari implicit subset, the high OOKB recall is bounded above by the candidate-recall ceiling (CR $= 0.67$), and the cases above that ceiling are OOKB predictions of this type.
- **Ranking errors with the correct candidate present.** For a non-trivial slice of mentions, the correct Wikidata QID is in the candidate set supplied to the ranker but the ranker selects a competitor. The within-artist same-location distractor cases used as few-shot demonstrations (Section 7.4.9.1) instantiate the discrimination challenge in this class: when several of a single artist's works share a location and medium, surface-form similarity carries little signal and the ranker must rely on iconographic or commissioner context. This failure mode contributes to KB-linked precision loss rather than to OOKB confusion.
- **Over-prediction of in-KB matches.** The ranker selects a candidate when no candidate in fact corresponds to the mention – the converse of the retrieval-miss case. The cases tend to occur when the candidate set contains a candidate whose surface form or descriptive fields overlap loosely with the mention but whose entity identity does not. ObliquER's ranker operates independently per mention (Section 7.4.9.4); each ambiguous mention is judged on its own context. A diagnostic instance is the linking of a Vasari artwork mention to `Q125760953` (*Fuente del Triunfo*), a fountain created in Córdoba in 1765, whose creation

date and place fall entirely outside the spatiotemporal envelope of *The Lives*. Neither the retrievers nor the ranker enforces a hard constraint on creation date or place, so a candidate that is impossible on these grounds but overlaps in a descriptive field can survive into the candidate set and be selected. The structured candidate records already carry inception (`wdt:P571`) and location (`wdt:P276`) fields (Table 7.13); these enter the ranker as soft descriptive context but are not imposed as admissibility filters. Turning them into explicit spatiotemporal constraints is a candidate-generation and ranking remediation taken up in Section 12.3.

The three failure modes occur at distinct stages of the pipeline (retrieval, ranker discrimination, and ranker precision-recall trade-off). The interpretive reading of each – the relation to the architectural choices of the retrieve-then-rank pipeline and the candidate remediations they suggest – is taken up in Section 11.3.

Parametric-only linking floor (A1 monolith)

The monolith ablation A1 (Section 7.4.8.4) collapses recognition, linking, and coreference into a single prompt and asks the LLM to produce both a typed mention span and a Wikidata identifier in one decoding pass, without any retrieval-stage support and without consulting the candidate-generation stack of Section 7.4.9.4. Per the methodology, A1 functions simultaneously as an upper bound on what the LLM can achieve at EL from parametric memory alone and as a lower bound on the contribution of the retrieve-then-rank architecture when added on top. A production system that adds retrieval must improve over A1, and the magnitude of that improvement bounds the retrieval contribution from below. A1's recognition-side scores under the same end-to-end run are reported separately in Table 9.17 of Section 9.4.6; the present subsection focuses on the linking-side scores of the same A1 run.

Table 9.15 reports the comparison. P/R/F1 are reported separately for KB-linked and OOKB mentions, mirroring Table 9.11. Candidate recall (CR) is not applicable to A1 because A1 does not consult a candidate set; it is reported for the two ObliquER rows as a reference point.

Table 9.15 reports an A1 KB-linked F1 of 0.000 across All, Explicit, and Implicit subsets, against ObliquER's 0.431 (gold mention) and 0.330 (end-to-end) on the All row – parametric memory alone links no mention correctly. Yet the A1 OOKB column is inflated, since its All-row OOKB F1 of 0.817 (P/R 0.715/0.952) exceeds the ObliquER gold-mention figure of 0.721 by 0.096 and the end-to-end figure of 0.750 by 0.067, with the Explicit-subset precision of 0.450 its lowest OOKB cell. The three rows use different mention bases – A1 and ObliquER end-to-end emit their own mentions, while the ObliquER gold-mention row evaluates on 1,010 manually annotated

Table 9.15: Linking floor (A1 monolith) vs. the production ObliquER linking pipeline on the Viewsari GT. A1 collapses recognition, linking, and coreference into one prompt and produces Wikidata identifiers from the LLM's parametric memory alone, without retrieval. CR is not applicable to A1 (no candidate set). The two ObliquER rows give the gold-mention and end-to-end settings, respectively, reproduced from Table 9.11 (*All* row) and Table 9.13 (*All* row); A1 is necessarily end-to-end because it emits its own mentions. The *n* column records the mention basis of each row. A1 evaluates on its own 1,422 emitted mentions (416 explicit + 1,006 implicit), the ObliquER end-to-end row on its own NER-emitted mentions, and the ObliquER gold-mention row on the 1,010 explicit-and-implicit mention instances of the GT. All A1 values extend the published evaluation.

		KB-linked			OOKB			
Configuration	**n**	**P**	**R**	**F1**	**P**	**R**	**F1**	**CR**
A1: Monolith prompt	1,422	0.000	0.000	0.000	0.715	0.952	0.817	–
A1, Explicit subset	416	0.000	0.000	0.000	0.450	0.806	0.578	–
A1, Implicit subset	1,006	0.000	0.000	0.000	0.805	0.986	0.886	–
ObliquER (end-to-end)	*NER-emitted*	0.280	0.400	0.330	0.820	0.690	0.750	0.590
ObliquER (gold mention, retrieve-then-rank)	**1,010**	**0.417**	**0.446**	**0.431**	**0.740**	0.703	0.721	**0.671**

mention instances – so the KB-linked-F1 comparison should be read across rows rather than as a single matched pairwise contrast.

The interpretive reading of Table 9.15 – the diagnostic value of the A1-vs-ObliquER gap, the symmetric reading against the recognition-side ablation of Table 9.17, the relationship between A1's zero KB-linked F1 and ReFinED's surface-form fallback pattern in Table 9.11, and what the inflated A1 OOKB column implies for parametric-only linking pipelines – is taken up in Section 11.3.

9.4.5 Cross-model evaluation

The numbers reported in Section 9.4.3 and Section 9.4.4 are all produced with a single inference model, `openai/gpt-oss-120b`, served from the GWDG *chat-AI* endpoint (see Section 7.4.8.1). A natural concern is that the observed performance is an artifact of that particular model rather than of the ontology-grounded prompt architecture. A different base model could plausibly close the gap to the discriminative baselines without any of ObliquER's structural commitments or, conversely, could reveal that the prompt design is brittle to the choice of decoder. This section addresses both possibilities by rerunning the recognition pipeline against a second instruction-tuned model, holding the prompt configuration constant, and reporting the resulting deltas on the Viewsari GT.

Protocol. The cross-model run holds every component of the pipeline fixed except the inference target. Concretely: (i) the GT corpus is the same 232-paragraph evaluation partition used in Table 9.6 after the unified few-shot exclusion set is removed (see the preamble of Section 9.4.6 and the dev-split description in Section 7.4.7); (ii) the recognition prompt is the ontology-guided variant with its two demonstrations from the Giotto (paragraph 112) and Ghirlandaio (paragraph 251) biographies, which is the configuration carried forward as the strongest recognition prompt in the ablation series of Section 9.4.6; (iii) the SemEval 2013-9.1 metrics, the per-mention-type confusion matrix, and the post-processing pipeline (label normalization, offset re-anchoring, cluster unification) are reused without modification; (iv) the sampling parameters (temperature, top-p, max-tokens) and the sliding-window chunking strategy are inherited from the `gpt-oss-120b` runs. The only change is the model identifier passed to the OpenAI-compatible *chat-AI* endpoint, which is switched from `openai/gpt-oss-120b` to `google/gemma-4-31b-it`. Both endpoints expose the same request schema, so no client-side changes are required, and the prompt rendering is bit-identical across models.

Isolated factors of the cross-model evaluation. Holding the prompt, the dev split, the metrics, and the post-processing constant means that the delta between the two models is attributable to two factors that vary together. Model capacity and pre-training distribution. We read the delta along three diagnostic axes rather than as a single number:

- the absolute F1 lift on the *Strict* criterion, which measures how much a stronger instruction-following model contributes on top of a fixed prompt;
- the relative lift on the *Type* vs. *Strict* criterion (the gap that already characterizes the `gpt-oss-120b` run in Table 9.7), which tests whether better decoding closes the boundary-error mode – i.e., whether mis-aligned spans on correctly typed mentions are an artifact of the smaller model or a property of the corpus;
- the implicit-vs-explicit ratio in the per-mention-type confusion matrix (Table 9.8), which tests whether the implicit-recovery surface that distinguishes ObliquER from the surface form-only baselines transfers to a different decoder.

Pre-registered expectation. The anticipation that `gemma-4-31b-it` would outperform `gpt-oss-120b` on the headline Strict F1, on the basis that the Gemma 4 family has stronger reported instruction-following on long-context structured-output tasks (Gemma Team, Google 2026) and that the prompt budget (entity definition + modeling example + two demonstrations + previous-paragraph context) places mild but consistent demands on long-context coherence that disadvantage the smaller of the two models. The more methodologically informative outcome was

the decomposition of F1 across mention types and SemEval criteria. A uniform lift across all three diagnostic axes would be consistent with the prompt-architecture hypothesis (the framework's contribution is invariant to model choice, and the headline numbers scale with model capability). A non-uniform lift – in particular, a disproportionate gain on the explicit-only subset combined with a flat implicit subset – would suggest that Gemma is better at the part of the task that surface-form-based baselines also solve, and that the implicit-recognition contribution of the ontology-typed prompt is closer to the smaller-model ceiling than the `gpt-oss-120b` number suggests.

Table 9.16: Cross-model recognition ablation on Viewsari, ontology-guided template held constant, evaluated on the updated Viewsari GT (Section 7.4.4). Recall (R) and F1 are reported under span-level bipartite matching across the Type, Partial, and Strict criteria for two inference models – `openai/gpt-oss-120b` and `google/gemma-4-31b-it` – at five demonstration counts $k \in \{0, 1, 2, 3, 5\}$. † marks the primary configuration used in the main results of Table 9.6. Both models served from the GWDG *chat-AI* endpoint with identical sampling parameters; *Precision* is omitted to keep the table compact (full P/R/F1 for the $k = 2$ rows is in Table 9.7 and Table 9.17).

	Type		Partial		Strict	
Shots	**R**	**F1**	**R**	**F1**	**R**	**F1**
openai/gpt-oss-120b						
$k = 0$	0.689	0.639	0.752	0.698	0.356	0.330
$k = 1$	0.721	0.611	0.789	0.668	0.362	0.307
$k = 2^{\dagger}$	0.751	0.642	0.815	0.697	0.391	0.334
$k = 3$	0.732	0.618	0.800	0.675	0.374	0.316
$k = 5$	0.731	0.612	0.816	0.684	0.410	0.344
google/gemma-4-31b-it						
$k = 0$	0.703	0.626	0.793	0.706	0.363	0.323
$k = 1$	0.741	0.606	0.827	0.676	0.410	0.336
$k = 2$	0.766	0.616	0.838	0.674	0.449	0.361
$k = 3$	0.762	0.624	0.834	0.683	0.437	0.358
$k = 5$	0.772	0.629	0.835	0.679	0.456	0.371

Observed outcomes. The three diagnostic axes set out above are populated as follows on Table 9.16.

Absolute headline lift. At the primary $k = 2$ configuration, Strict F1 rises by 0.027 from `gpt-oss-120b` to `gemma-4-31b-it` ($0.334 \rightarrow 0.361$); Type F1 falls by 0.026 ($0.642 \rightarrow 0.616$); Partial F1 falls by 0.023 ($0.697 \rightarrow 0.674$). Neither decoder dominates across the three criteria. The leading model flips by criterion and by shot count – `gemma-4-31b-it` leads on Strict F1 at

every shot count except $k = 0$, the two are mixed on Partial, and `gpt-oss-120b` leads on Type F1 at $k \leq 2$ with `gemma-4-31b-it` overtaking at $k \in \{3, 5\}$ – which is the close cross-model proximity the section tests for rather than a decisive capability difference. The A3-equivalent zero-shot row ($k = 0$) returns Strict F1 $= 0.330$ on `gpt-oss-120b` and 0.323 on `gemma-4-31b-it`, a gap of 0.007 – the tightest cross-model agreement in the table. Across all ten (k, model) cells the Strict F1 spans a narrow 0.064 band (0.307–0.371), the close envelope the section tests for.

Type-minus-Strict gap. The gap is preserved across both models at comparable magnitude, although consistently smaller under `gemma-4-31b-it`. At $k = 2$, `gpt-oss-120b` reports a gap of 0.308 $(0.642 - 0.334)$ between Type and Strict F1; `gemma-4-31b-it` reports 0.255 $(0.616 - 0.361)$. The gap is neither model-specific nor closed by the larger decoder.

Few-shot count axis. Under `gpt-oss-120b`, F1 is largely stable across $k \in \{0, 1, 2, 3, 5\}$ as recall rises and precision falls in step. Under `gemma-4-31b-it`, the recall trend with more examples is more monotone – Type R rises from 0.703 ($k = 0$) to 0.772 ($k = 5$), Strict R from 0.363 to 0.456 – and Strict and Type F1 rise with k (Partial F1 instead peaks at $k = 0$ and drifts down), with $k = 5$ topping the Strict and Type columns under Gemma. The $k = 2$ configuration is not strictly optimal under either model in isolation, but it sits on a stable plateau where moving along the shot axis does not shift F1 dramatically in either direction.

The interpretive reading of these three readings against the prompt-architecture hypothesis – the bearing of the cross-model stability on whether the framework's contribution is invariant to model choice; the role of the close $k = 0$ Strict F1 agreement (0.330 vs. 0.323) in attributing the zero-shot recognition surface to the ontology layer rather than to the decoder; the relationship between the preserved Type-minus-Strict gap and the boundary-detection failure mode of Section 9.4.3; and the consequences for the production $k = 2$ choice – is taken up in Section 11.3.

9.4.6 Ablation study

To decompose the contribution of the individual architectural layers of ObliquER, we report four ablation conditions, each targeting a distinct design decision of the pipeline. Each condition corresponds to a self-contained prompt variant, configured so that switching between them requires only a single Hydra override at the inference entry point (Section 7.4.8.1). The ObliquER reference row is the production ontology-guided k=2 configuration, reported under span-level bipartite matching in Table 9.7 and Table 9.16, with its per-mention-type behavior in the confusion matrix of Table 9.8.

A note on the evaluation split is required upfront. Because A4 varies the number of in-context demonstrations from one to five, the five demonstration paragraphs used in the largest few-shot

configuration (Ghirlandaio 251, Giotto 112, Titian 250, Brunelleschi 206, and Lippi 9) must be excluded from the test set to avoid prompt/test leakage, together with the Cimabue 1 paragraph that supplies the worked modeling example carried in the ontology-guided prompt. The 5-shot demonstration set is a strict superset of the 3-shot and 1-shot sets, so all configurations are evaluated on a single shared test partition that excludes the union of these six paragraphs; of the 1,010 implicit-and-explicit gold mentions, the 30 that fall in these paragraphs are removed, giving the 980-mention recognition partition of Section 9.4.3. The unified exclusion set is applied uniformly across A1–A4, ensuring the resulting numbers are directly comparable.

Ablation A1: Monolith prompt. The three-stage sub-tasks of the ObliquER pipeline – entity recognition, EL, and coreference resolution – are collapsed into a single prompt. The model is asked, in one pass, to (i) identify all artwork mentions (explicit, implicit, coreferent, generic), (ii) link each recognized mention to a Wikidata Q-identifier drawn from its own pre-training knowledge, without access to the dense-retrieval candidate-generation stage described in Section 7.4.9.4, and (iii) resolve coreference chains within and across paragraph boundaries. The output schema extends the standard recognition record (Section 7.4.6.2) with an additional Wikidata-identifier field per mention; all other fields remain unchanged, so that recognition metrics are computed against the same GT and under the same SemEval 2013-9.1 span-level criteria as the baseline comparison in Table 9.6.

What this tests. Whether the staged pipeline (chunking $\rightarrow$ NER $\rightarrow$ EL $\rightarrow$ COREF) yields measurable gains over a single end-to-end prompt, and specifically whether the separation of recognition from linking is empirically justified or imposes unnecessary pipeline overhead.

Ablation A2: No domain-specific content. The modular prompt structure is preserved (identical output schema, identical mention-type vocabulary, identical prior-output slot for cross-paragraph context), but all domain-specific content is removed from the five prompt blocks described in Section 7.4.8.3. Concretely: (i) the Renaissance art-historical expert role is replaced with a generic named-entity-recognition assistant role; (ii) the ontology-grounded entity definition – including the enumeration of artwork subtypes (paintings, frescoes, sculptures, architectural works) and the explicit exclusion of process descriptions such as *skill in fresco* – is replaced with a minimal one-sentence definition (*"an artwork is a creative product"*); (iii) the special-cases block addressing nested mentions, generic collections, and buildings-as-artworks is removed; and (iv) the few-shot example block is rendered empty, reducing the prompt to a generic zero-shot NER instruction.

What this tests. Whether the domain-specific content layer – expert framing, artwork taxonomy, edge-case rules, and few-shot demonstrations – contributes measurably to extraction quality,

independently of the pipeline structure (which is kept identical to the full ObliquER configuration). A1 and A2 are complementary. A1 strips the staging but keeps the content, A2 strips the content but keeps the staging, and the gap between them quantifies which of the two contributes more.

Ablation A3: Zero-shot. The ontology-guided prompt structure is preserved in full – the expert role, the ontology-typed entity definition (using the four mention subclass names: *explicit artwork mention*, *implicit artwork mention*, *coreferent*, *generic mention*), the special-cases block, and the modeling-example block that maps an extracted mention to its Viewsari KG serialization – but the demonstration list is empty, so the model receives zero in-context demonstrations of annotated paragraphs and proceeds directly from the structural blocks to the current paragraph.

What this tests. A3 isolates the contribution of in-context demonstrations from the rest of the ontology-guided prompt. The ontology-guided variant (which the cross-model evaluation of Section 9.4.5 carries forward as the strongest recognition configuration) packages two distinct knowledge sources: (i) the ontology-typed entity definitions and the modeling-example block, which align the model's vocabulary and output structure with the Viewsari schema, and (ii) the two annotated demonstrations (Giotto paragraph 112 and Ghirlandaio paragraph 251), which illustrate the application of that schema to a complete paragraph. A2 strips both layers; A3 strips only the demonstrations, leaving the ontology layer intact. The difference between A2 and A3 therefore attributes the recognition gain to the ontology-typed definitions independently of any specific demonstration, while the difference between A3 and the full ontology-guided configuration (the $k = 2$ row of Table 9.18) attributes it to the demonstrations.

Ablation A4: Few-shot count. A4 holds the ontology-guided template fixed (same structural blocks as A3 and as the production configuration) and varies only the length of the demonstration list, with $k \in \{1, 3, 5\}$ demonstrations. The example sets are nested: $k = 1$ uses Ghirlandaio paragraph 251 (12 mentions; mixed mention types, nested spans, multi-hop coreference chains); $k = 3$ adds Giotto paragraph 112 (11 mentions; all four mention types including *generic mention*, mostly Wikidata-linkable) and Titian paragraph 250 (12 mentions; long coreference chains, predominantly implicit references); $k = 5$ further adds Brunelleschi paragraph 206 (10 mentions, 8/10 coreferent – a pure coreference stress case) and Lippi paragraph 9 (8 mentions; balanced mix across all four types in a single paragraph). The five examples are chosen to expose distinct phenomena observed during prompt development (Section 7.4.8.3); each is drawn verbatim from the GT corpus and is excluded from the test partition under the unified exclusion set described above.

What this tests. Whether the observed pipeline performance is sensitive to the number of in-context demonstrations or whether the ontology-guided template plateaus quickly with respect

to shot count. The $k = 2$ configuration – which is the one used by the production ontology-guided prompt and carried forward to the cross-model evaluation – is intentionally absent from A4 because it is reported in the reference rows of Table 9.18; this lets A4 trace the curve through $k \in \{0, 1, 2, 3, 5\}$, where the $k = 0$ value reuses the A3 row. If the curve flattens between $k = 2$ and $k = 5$, the production choice sits on a plateau and is robust to small variations in demonstration budget; if the $k = 5$ value is materially higher, the production prompt under-uses the demonstration channel and the cross-model results of Section 9.4.5 are conservative; if $k = 5$ degrades relative to $k \in \{2, 3\}$, the demonstration block is competing with the ontology blocks for context budget and additional demonstrations dilute rather than reinforce the schema.

Expected outcomes (A1–A3). Table 9.17 reports A1–A3 against the full ObliquER configuration. The monolith prompt (A1) was expected to degrade EL F1 most substantially – since it lacks access to artist-constrained candidate retrieval – while degrading recognition F1 modestly, as the model retains the same domain context but must coordinate three sub-tasks in a single decoding pass; this pattern would reflect the limitation identified by Wang et al. (2025a) regarding concurrent structured operations in generative decoding (Section 5.2.5). The removal of domain content (A2) was expected to degrade recognition precision the most, since the entity definition block is the primary mechanism for suppressing false positives from generic artistic references. The zero-shot condition (A3) was expected to fall between A2 and the full ontology-guided configuration. The ontology-typed definitions provide most of the schema alignment, but the demonstrations were predicted to yield a non-trivial increment in implicit mentions, whose recognition depends on a pattern more easily conveyed by example than by definition.

Observed outcomes (A4). Table 9.18 reports A4 with the recall and F1 values from the updated Viewsari GT. The pre-registered expectation of a monotone non-decreasing curve from $k = 0$ to $k = 2$ is *not* borne out. Strict F1 dips at $k = 1$ (0.307) below the zero-shot baseline ($k = 0$: 0.330), recovers at $k = 2$ (0.334), dips again at $k = 3$ (0.316), and reaches its top at $k = 5$ (0.344). The Partial column tracks a similar oscillation in a narrower band (0.668 – 0.698). Across $k \in \{0, 1, 2, 3, 5\}$ the Strict F1 envelope is 0.037 (min 0.307 at $k = 1$, max 0.344 at $k = 5$); the $k = 5$ value exceeds $k = 2$ by 0.010 on Strict F1, with $k = 2$ retaining a near-peak Partial F1 (0.697 vs. 0.684 at $k = 5$). Under `gemma-4-31b-it` (cross-model rows of Table 9.16), the same non-monotone pattern repeats with the global maximum again near $k \in \{2, 5\}$.

Observed outcomes (A1, A2, A3). The three architectural ablations populate the A1–A3 rows of Table 9.17 under both inference models and resolve the *Expected outcomes (A1–A3)* above.

Table 9.17: Architectural ablation (A1–A3) on the Viewsari GT under span-level bipartite matching, reported under both inference models `openai/gpt-oss-120b` and `google/gemma-4-31b-it` (Strict P/R/F1 plus headline F1 under Partial and Type). The ObliquER reference rows reproduce the corresponding $k = 2$ entries of Table 9.16.

		Strict			F1 by criterion	
Configuration	**Model**	**P**	**R**	**F1**	**Partial**	**Type**
A1: Monolith prompt	`gpt-oss-120b`	0.268	0.301	0.284	0.671	0.577
	`gemma-4-31b-it`	0.257	0.360	0.300	0.684	0.609
A2: No domain-specific content	`gpt-oss-120b`	0.269	0.233	0.249	0.581	0.526
	`gemma-4-31b-it`	0.241	0.209	0.224	0.591	0.512
A3: Zero-shot (ontology, no demos)	`gpt-oss-120b`	0.308	0.356	0.330	0.698	0.639
	`gemma-4-31b-it`	0.291	0.363	0.323	0.706	0.626
ObliquER (prod., k=2, ref.)	`gpt-oss-120b`	**0.292**	**0.391**	**0.334**	**0.697**	**0.642**
	`gemma-4-31b-it`	**0.302**	**0.449**	**0.361**	**0.674**	**0.616**

A1 (monolith prompt) returns Strict F1 $= 0.284$ on `gpt-oss-120b` against the reference 0.334 ($\Delta = -0.050$, a 15 % relative drop) and Strict F1 $= 0.300$ on `gemma-4-31b-it` against the reference 0.361 ($\Delta = -0.061$, a 17 % relative drop); the Type F1 drops are -0.065 (GPT) and -0.007 (Gemma). The recognition-F1 portion of the pre-registered prediction – A1 "degrading recognition F1 modestly" – is borne out under both decoders, and the staging gap is comparable in magnitude across models (0.050–0.061 Strict F1). The linking-side prediction is addressed by Section 9.4.4.

A2 (no domain-specific content) is the harshest of the architectural ablations on both decoders. Strict F1 drops to 0.249 on GPT ($\Delta = -0.085$ vs. the GPT reference, 25 % relative drop) and to 0.224 on Gemma ($\Delta = -0.137$ vs. the Gemma reference, 38 % relative drop), with comparable drops on Type F1 (-0.116 GPT, -0.104 Gemma) and Partial F1 (-0.116 GPT, -0.083 Gemma). The pre-registration anticipated a precision-driven failure mode; the data show Strict P falling from 0.292 at the GPT reference to 0.269, and from 0.302 at the Gemma reference to 0.241, alongside a Strict R drop of similar magnitude under each decoder that the pre-registration had not named. The Strict, Type, and Partial drops are of comparable magnitude in both models.

A3 (zero-shot ontology, no demos) lands close to the reference under both decoders. Strict F1 $= 0.330$ on GPT (against 0.334, $\Delta = -0.004$) and Strict F1 $= 0.323$ on Gemma (against 0.361, $\Delta = -0.038$). On GPT, the demonstration contribution is therefore essentially zero on Strict F1; on Gemma, it is more substantial. The position of A3 "between A2 and the full ontology-guided configuration" on the assumption that the two annotated demonstrations

would "carry a non-trivial increment on implicit mentions"; the prediction is contradicted under GPT (the GPT $\Delta = -0.004$ sits well within the bootstrap intervals of Table 9.5, where A3 GPT is $[0.299, 0.361]$ and the production reference is $[0.300, 0.368]$) and qualifiedly borne out under Gemma (the Gemma $\Delta = -0.038$ is comparable to the bootstrap-interval half-width and represents the largest single-decoder demonstration contribution observed across the chapter).

Decomposition of the content-layer contribution. Reading A2, A3, and the reference together decomposes the content-layer contribution between its two sub-components, separately under each decoder. Under `gpt-oss-120b`. Total content-layer contribution (reference − A2) = $+0.085$ Strict F1; the ontology layer alone (A3 − A2) = $+0.081$; the demonstration layer (reference − A3) = $+0.004$. The ontology-typed entity definitions, the special-cases block, and the KG modeling example therefore account for approximately 95 % of the GPT content-layer contribution, with the demonstrations accounting for the remaining 5 %. Under `gemma-4-31b-it`. Total content-layer contribution = $+0.137$; ontology layer = $+0.099$ (72 %); demonstration layer = $+0.038$ (28 %). The split inverts the assumption on both decoders – the demonstrations were predicted to carry a non-trivial increment on implicit mentions – and the inversion is sharper on GPT than on Gemma, indicating that Gemma uses the demonstration channel more than GPT but that the ontology layer dominates the content-layer contribution under either decoder.

The interpretive reading of A1, A2, A3 together – the rank-ordering of architectural pillars (ontology layer dominant under both decoders, pipeline staging contributing 0.050–0.061 Strict F1, demonstrations near-zero under GPT and modest under Gemma), the bearing of the 95 / 5 (GPT) and 72 / 28 (Gemma) decompositions our previous assumption, the re-framing of the few-shot demonstrations, and the joint reading against the cross-model evidence of Table 9.16 – is taken up in Section 11.3.

Model scope of the architectural ablations. A1–A3 are reported under both `openai/gpt-oss-120b` and `google/gemma-4-31b-it` in Table 9.17, and the qualitative rank-ordering of the three pillars – ontology-typed content layer above pipeline staging above demonstrations on Strict F1 – holds under each decoder; the weight of the demonstration layer is the one quantitative difference (essentially zero under GPT, $+0.038$ under Gemma).

Prompt-variant comparison: few-shot, ontology-guided, chain-of-thought

The methodology of Section 7.4.8.3 describes three production-grade recognition prompt variants. The *few-shot* baseline, which embeds an artwork description without ontological grounding; the *ontology-guided* variant, which carries the full ontology-typed entity definition, the special-cases

Table 9.18: Few-shot count ablation (A4) on the Viewsari GT under span-level bipartite matching, ontology-guided template held constant for `gpt-oss-120b` (Section 7.4.4). *Strict* and *Partial* are the span-level criteria of Table 9.7. The $k = 0$ row matches the GPT A3 row of Table 9.17; the $k = 2$ row is the GPT production reference. The demonstration-paragraph column is cumulative. Each entry lists the paragraphs added relative to the row above. The corresponding `gemma-4-31b-it` curve is reported in Table 9.16. All rows are evaluated on the unified-exclusion test partition described in the preamble of this subsection.

		Strict		Partial	
Variant	**Demo paragraphs**	**R**	**F1**	**R**	**F1**
$k = 0$ (= A3)	–	0.356	0.330	0.752	0.698
$k = 1$	Ghirlandaio 251	0.362	0.307	0.789	0.668
$k = 2$ (prod.)	+ Giotto 112	0.391	0.334	0.815	0.697
$k = 3$	+ Titian 250	0.374	0.316	0.800	0.675
$k = 5$	+ Brunelleschi 206, Lippi 9	0.410	0.344	0.816	0.684

block, and the KG modeling example, and is the configuration carried forward as the production reference in Table 9.17 and Table 9.18; and the CoT variant, which adds a structured reasoning trace. The architectural ablation of Table 9.17 compared the ontology-guided variant against its content-stripped (A2) and demonstration-stripped (A3) descendants, but did not compare against the alternative full prompt structures. Table 9.19 fills that gap under `openai/gpt-oss-120b`, holding the corpus, the evaluation split, the sampling parameters, and the chunking strategy constant across the three variants and reporting the standard SemEval suite plus the surface-form Strict and Partial columns. The few-shot variant uses a single demonstration paragraph drawn from the unified-exclusion set so that the variant comparison sits on the same evaluation partition as the rest of this section.

Table 9.19: Prompt-variant comparison on the Viewsari GT under `openai/gpt-oss-120b` and span-level bipartite matching, evaluated on the unified-exclusion test partition. Each row varies the prompt structure while holding the chunking strategy, the evaluation suite, and the sampling parameters constant. The Ontology-guided ($k = 2$) row reproduces the GPT production reference of Table 9.17.

	Strict			F1 by criterion	
Prompt variant	**P**	**R**	**F1**	**Partial**	**Type**
Few-shot baseline	0.281	0.346	0.310	0.658	0.587
Ontology-guided ($k = 2$, prod.)	**0.292**	**0.391**	**0.334**	**0.697**	**0.642**
CoT	0.295	0.366	0.327	0.684	0.609

Observed outcomes (prompt-variant comparison). The ontology-guided variant tops every column in Table 9.19. The Strict F1 spread across the three variants is 0.310 (Few-shot), 0.327 (CoT), 0.334 (Ontology-guided), and a 0.024 envelope. The Type F1 spread is wider, at 0.587 (Few-shot), 0.609 (CoT), 0.642 (Ontology-guided), and a 0.055 envelope, the largest single-criterion gap among the three variants. The Few-shot baseline trails the ontology-guided variant by 0.024 Strict F1 and 0.055 Type F1, consistent with the architectural-ablation reading that the ontology-typed entity definitions and the KG modeling example carry the bulk of the content-layer contribution. The CoT variant sits between the two, improving on the Few-shot baseline by 0.017 Strict F1 but not matching the Ontology-guided F1 on any of the three criteria. The pre-registered open empirical question stated in Chapter 7 – whether CoT consistently improves extraction accuracy over the ontology-guided variant – resolves negatively under `openai/gpt-oss-120b`. The CoT variant pays a small accuracy cost relative to the production ontology-guided template, while delivering the structured reasoning trace as a provenance artifact. The interpretive reading of these three variants against the prompt-architecture hypothesis, in particular the asymmetry between the recognition-F1 cost of the CoT variant and its provenance-mechanism value (Section 9.3), is taken up in Section 11.3.

Reproducibility. Each ablation condition is selected via a single Hydra override that switches the dataset configuration at the recognition entry point; the unified evaluation exclusion list is applied identically across all configurations, ensuring that the GT filtering does not depend on the prompt variant in use. The exact configuration names, prompt directories, and invocations are documented in the project repository.

9.4.7 ArtPedia. Transferability assessment

The evaluation on ArtPedia assesses the transferability of ObliquER to a linguistically modern, domain-analogous corpus. The experimental setup is described in Section 7.4.10.1; the headline figures are reproduced in Table 9.6 and Table 9.7.

ArtPedia's structural constraints – sentence-level surface forms, one artwork per paragraph – differ from the Viewsari annotation schema, and the evaluation protocol is adapted accordingly. The analysis is restricted to explicit and implicit mention types, and precision is treated as informative only for the single annotated artwork per entry. A direct F1 comparison between ArtPedia and Viewsari results is therefore not made; the comparison reported here is directional, addressing the implicit-mention recall that ObliquER achieves in a domain it was not designed for, under the same production prompt configuration.

Table 9.6 reports ObliquER on ArtPedia at a surface-form Strict F1 $= 0.319$ (P $= 0.330$, R $= 0.308$) and Partial F1 $= 0.472$ (P $= 0.489$, R $= 0.457$), against 0.334 and 0.494 respectively on Viewsari. The surface-form Strict-F1 drop from Viewsari to ArtPedia is thus $0.334 \rightarrow 0.319$ (-0.015); the span-level decomposition of Table 9.7 gives the corresponding drops as Strict $0.334 \rightarrow 0.321$, Type $0.642 \rightarrow 0.587$, and Partial $0.697 \rightarrow 0.608$. Under the same constraints, the surface form-only baselines return Strict F1 figures of 0.09 (UniNER-7B) and 0.09 (GLiNER2) on ArtPedia; the corresponding ObliquER row reads 0.319 (Table 9.6).

The confusion matrix on ArtPedia (Table 9.9) shows that the dominant error direction is the misclassification of descriptive implicit references as explicit ones. The direction is the same as on Viewsari (implicit $\rightarrow$ *Explicit* dominates at 57 cases against 14 in the reverse direction, out of 771 implicit gold mentions), magnified by ArtPedia's order-of-magnitude larger implicit subset (19,549 implicit gold mentions vs. 771 on Viewsari). The principal recall failure on ArtPedia is the 8,572 *missed* implicit mentions ($\approx 44\,\%$ of the implicit subset) – a coverage failure that does not have a Viewsari analog at this magnitude and that reflects the sheer scale and density of implicit descriptive references in ArtPedia's Wikipedia-grounded paragraphs.

The interpretive reading of the ArtPedia transfer – the attribution of the cross-corpus F1 drop to corpus properties rather than pipeline failure, the bearing of the ObliquER-vs-baseline gap on the prompt-architecture hypothesis under a period and register shift, and what the same-direction-magnified error pattern implies about the descriptive-paraphrase / canonical-title boundary as a corpus-level rather than model-level phenomenon – is taken up in Section 11.3. Scope limitations that bear on the ArtPedia transfer (single-language Viewsari corpus, deferred cross-version comparison, motif and iconographic reconciliation as future capability) are reported alongside the other limitations of the present work in Section 11.9.

9.5 Knowledge graph: measuring the population results

All metrics in this section are computed over population Run 2, the current build of the populated graph defined in Section 8.3; figures are reported against that artifact, with the earlier Run 1 retained only as a documented prior state against which the graph's growth is described (Section 8.3.5). The graph is a living artifact whose full metric suite is re-run at each regeneration, so the figures below are a snapshot of the current build rather than a fixed, once-and-for-all measurement. The recognition and linking evaluation, by contrast, continues to use the corrected GT corpus that the KG does not yet incorporate (Section 7.4.7), so the two families of figures rest on deliberately different GT versions.

The headline counts of the populated Viewsari KG – 2,455,976 triples across 112,113 annotations, 86,654 typed mentions, 15,804 artwork entities, 443 person entities, 632 co-occurrences, 13,794 extraction activities, and 165 biographies across 10 volumes – are reported as descriptive corpus statistics in Section 8.3 of the Results chapter and are not re-tabulated here. The present section reports the *evaluative* reading of those numbers in light of three quantitative claims that the methodology stages as falsifiable. The upper bound on verbosity, the distribution of external links, and the bibliographic reachability of the FRBR-aligned chain. The two remaining evaluative metrics, provenance coverage M06 and strict provenance coverage M07, are reported alongside the ontology's provenance design in Section 9.5.4 of the ontology evaluation, since they address the ontology's structural commitments rather than the population outcome. Table 9.20 catalogs the five evaluative metrics, the section in which each is reported, and its pass criterion. The interpretive reading of the three KG-population metrics against the dissertation's central claim and C1–C4 is taken up in Section 11.2.

Table 9.20: Catalog of evaluative KG metrics, the section in which each is reported, and its pass criterion. Descriptive counterparts (M01–M04, M08, M10, M11) are reported in Section 8.3.

ID	Metric	Reported in	Pass criterion
M05	Triples per mention (verbosity proxy)	Section 9.5.1	Ratio $\leq$ 18 (CIDOC-comparison upper bound)
M06	Provenance coverage (loose)	Section 9.5.4	Ratio $= 1.0$
M07	Provenance coverage (strict, para+act+agent)	Section 9.5.4	Ratio $<$ M06; residual = human-agent + non-mention nodes
M09	External linking coverage by entity type	Section 9.5.2	Person rate $\gg$ artwork rate; long-tail
M12	Web-manifestation reachability	Section 9.5.3	Volume & biography 100%; page bounded

9.5.1 Verbosity: Triples per mention

The CIDOC comparison (Section 9.2.2) noted that the Viewsari representation pays for its provenance transparency with additional triples per mention, deriving the 12–18-triple range as the span between an unlinked and a fully resolved mention. Metric M05 measures the realized verbosity on the closed KG. 839,484 triples incident on a mention node (as subject or object) divided across 86,654 mention instances, for a mean of **9.7** triples per mention. The pass criterion (closed-graph ratio at or below the 18-triple ceiling) is met with substantial margin.

Two qualifications bear on the reading of the M05 ratio. The bulk of the per-mention triple budget is consumed by the annotation, target, body, and provenance fields that the Web Annotation

Ontology and PROV-O integration require by construction; the Viewsari-specific vocabulary contributes a smaller marginal cost, so the realized ratio is dominated by upstream commitments rather than by Viewsari's own modeling decisions. The ratio is computed over the full closed graph, including the 7,431 closure triples materialized by the rules in Listing A.2; the asserted-graph figure is correspondingly lower.

9.5.2 External linking coverage by entity type

Metric M09 measures the proportion of person and artwork entities reconciled to Wikidata via `owl:sameAs` (Table 9.21). The two principal entity classes diverge sharply in linkage rate: 320 / 443 persons (**72.2 %**) are reconciled to Wikidata identifiers, against 3,275 / 15,804 artworks (**20.7 %**); the resulting **12,529 OOKB artworks** retained under stable Viewsari-internal IRIs constitute 79.3 % of the artwork population. The pass criterion of M09 (person rate substantially exceeds artwork rate; artwork shortfall in the direction predicted by the long-tail discovery argument of Section 7.4.9.4) is met, and the 51.5-percentage-point gap is large and monotone in the expected direction, and it widens relative to Run 1 (a 35.7-point gap at 36.5 % artwork coverage) as the resolved scope expands from the stratified GT to the full recognition layer and recovers a larger long tail of works absent from Wikidata.

Table 9.21: External linking coverage by entity type on the populated KG (metric M09).

Entity class	Total	Wikidata-linked	Linked %
Person	443	320	72.2
Artwork	15,804	3,275	20.7

The two authority namespaces reached through `owl:sameAs` on the populated graph (M11, descriptive) are both Wikidata variants (`https://www.wikidata.org`: 3,164 links; `http://www.wikidata.org`: 431 links, for 3,595 in total – the sum of the 320 linked persons and 3,275 linked artworks of Table 9.21); no GeoNames or Iconclass links have been materialized in the current population pass, consistent with the locations- and motifs-as-future-work scope of Section 11.9.

9.5.3 Bibliographic reachability

The Viewsari bibliographic chain encodes every paragraph as a member of a biography, volume, translation, and edition (Section 7.3.1). Metric M12 measures whether every expression-level

node in this chain – volume, biography, page – resolves to a Project Gutenberg web manifestation via `frbr:hasEmbodiment` and `rdfs:seeAlso`. The result (Table 9.22) is that all 10 volumes and all 165 biographies reach a web URL (100 % coverage); page-level reachability lands at 2,794 / 3,026 pages (**92.3 %**). The 7.7 % page-level shortfall (232 pages) is bounded by the upstream HTML's anchor coverage. Pages whose Project Gutenberg manifestation does not expose a per-page anchor cannot be resolved by automated lookup against the existing FRBR chain. The pass criterion (volume and biography at 100 %; page-level reachability bounded by upstream HTML anchors) is met.

Table 9.22: Web-manifestation reachability per expression-level node (metric M12).

Expression level	Total	With web URL	%
Volume	10	10	100.0
Biography	165	165	100.0
Page	3026	2794	92.3

9.5.4 Provenance transparency

The provenance transparency evaluation quantifies the degree to which the Viewsari KG's annotation triples are traceable to their computational origin.

The result of the SPARQL provenance coverage query for the closed KG (Listing A.3), is reported here in the context of the methodological claim it supports. On the closed graph, the `withProv/total` ratio evaluates to 112,113 / 112,113 = **1.0**. Every `oa:Annotation` is linked through a `prov:wasGeneratedBy` chain to a `prov:Activity` that is in turn `prov:wasAssociatedWith` a documented `prov:SoftwareAgent`. This 100% provenance coverage result contrasts with the baseline KG of Santini et al. (2022, 2024), in which 0% of co-occurrence edges carry any form of computational provenance. In this case, zero provenance coverage means that no query over the baseline KG of that approach can determine which agent produced a given co-occurrence edge, under which configuration, or from which paragraph. 100 % provenance coverage in Viewsari means that every assertion is epistemically attributable. A researcher can determine, for any triple in the graph, the paragraph that sourced it, the ObliquER run that produced it, and the model version and prompt template used to produce the interpretation (or scripts for co-occurrences).

The 1.0 ratio is the joint product of two design decisions – the ontology defines the provenance edge, and every annotation ObliquER generates populates the required chain.

The second provenance metric concerns the proportion of annotations that are simultaneously traceable to paragraph, activity, and software agent; this is essentially a stricter criterion than traceability to any one of these. The strict ratio (metric M07) evaluates to 84,215 / 112,113 = 0.751 (75.1%) on the closed graph, the 84,215 being exactly the ObliquER-produced subset of the 86,654 typed mention annotations (Section 8.3.5). Every `oa:Annotation` that simultaneously reaches a paragraph, a typed activity, and a software agent through its provenance chain. The 24.9 % shortfall is not a coreference-anchoring artifact but a consequence of the criterion's `prov:SoftwareAgent` requirement. Of the 27,898 annotations outside the strict set, 2,439 are the manually annotated GT mentions, which carry a direct *in-paragraph* edge and a full activity chain but are attributed to a human annotator (`prov:Person`/`foaf:Person`) rather than to a `prov:SoftwareAgent`; the remaining 25,459 are non-mention annotation nodes (typed `oa:Annotation`/`prov:Entity`) that carry neither a direct *in-paragraph* edge nor a software-agent activity. This residual is all but identical to Run 1's 2,439 + 25,469 = 27,908 annotations, the difference being the 10 self-pair co-occurrence annotations dropped in the co-occurrence-layer correction; the ratio rises from Run 1's 0.664 purely because the machine-extracted numerator grew from 55,246 to 84,215 with the third recognition variant. Every typed mention – including all 22,283 coreferents – does carry the direct edge, so the strict ratio measures the machine-extracted, paragraph-anchored fraction specifically. The human-annotated gold layer is excluded by the agent-type condition rather than being missing from the graph.[8]

The interpretive reading of this contrast is taken up in Section 11.2.

9.5.5 Methodological claims supported by the knowledge graph population

The three evaluative metrics M05, M09, and M12, combined with the 100 % provenance coverage of Section 9.5.4, support the verbosity-as-affordability claim against the CIDOC CRM comparison. Additionally, it underlines the OOKB-as-long-tail-work claim against the surface-form-based linking regime and the FRBR-chain-materialized-end-to-end claim against a paper-commitment counterfactual. The descriptive corpus statistics of Section 8.3 provide the population substrate against which these evaluative readings sit; the interpretive reading of the three claims against the

[8] The 2,439 GT mentions counted here are the *pre-correction* figure. The KG population metrics (M01–M12) are computed over the deployed graph, whose GT layer was ingested before the headless-coreferent correction, which reduced the annotated corpus to 2,422 mentions (Section 7.4.7), and reconciliation is deferred to the next redeployment. The recognition and linking evaluations of this chapter instead use the corrected corpus (the 1,010 implicit-and-explicit mentions and their unified-exclusion partition), so the two families of figures are computed on different GT versions by design.

dissertation's central claim and C1–C4 is taken up in Section 11.2; the current scope limitations are contextualized in Section 11.9.

10 Beyond Viewsari: a transferability study on another source

In the preceding chapters, we have demonstrated the Viewsari methodology on the primary case study *The Lives* and evaluated it in Chapter 9, including a controlled art-historical corpus shift to ArtPedia (Section 9.4.7). This chapter is deliberately *exploratory*, extending the evaluation to examine the transferability of both the ontological framework and the *ObliquER* extraction pipeline *beyond* the primary domain. We chose a structurally distinct corpus to address **RQ3** as stated in Section 1.3.[1]

> **RQ3**: What reusable ontology design pattern emerges from modeling extraction as a documented, provenance-aware interpretive process, and *how* does that foundation transfer to structurally distinct (interpretive) corpora?

We aim to answer RQ3 with **E4**, which constitutes the transferability pilot to Georg Burckhard Kloß' *Annalen der Loge zur Einigkeit* (Section 10.2), linked to FactGrid (Simons 2024),[2] and **C4**, the *Computational Provenance* (COMP-PROV) design pattern, which was abstracted from an analysis of related ontologies in the digital humanities (Section 10.3), building on our work from Ondraszek et al. (2024c).

The transferability claim rests on two methodologically distinct axes. Section 9.4.7 provides the empirical axis, holding the extraction and linking task constant and shifting only the corpus to ArtPedia, a period- and register-shifted English-language collection with the same Wikidata target; this yields quantitative transfer evidence in the form of KB-linked F1 scores directly comparable to the Viewsari baseline. The Kloß/FactGrid pilot of Section 10.2 provides the conceptual axis: it simultaneously shifts language, century, entity type, target knowledge base, and referential strategy, demonstrating that the Viewsari ontology and COMP-PROV transfer

[1] Parts of this chapter build on previously published work: the comparative ontology-pattern analysis in Ondraszek et al. (2024c), updated and extended in Section 10.3, and the NFDI4Memory infrastructure that provided the context for the transferability pilot in Ondraszek et al. (2025a,b).

[2] The materials for this transferability pilot — the plain-text transcriptions, the FactGrid-derived GT, the adapted prompts, and a demonstration Viewsari RDF serialization of the GT — are versioned in the ObliquER repository: `https://github.com/ISE-FIZKarlsruhe/ObliquER/tree/main/data/factgrid`.

through the prompt content layer alone without end-to-end entity linking against FactGrid, whose evaluation is deferred to future work (Section 12.3). The pilot's contribution is conceptual rather than metric, showing that human-curated and machine-extracted provenance coexist in the same graph under the same schema, which is the claim a Yin-style theoretical replication is designed to establish (Yin 2018).

10.1 FactGrid and the Illuminati order of Frankfurt as a controlled transferability pilot

This pilot constitutes a theoretical replication in the sense of Yin (2018), functioning as a demonstration that the methodology can be systematically reapplied to a case that diverges on key dimensions while maintaining its core design commitments (Wieringa 2014d). The Viewsari ontology core, the ObliquER pipeline templates, the post-processing procedure, and the provenance overlay remain fixed; only the prompt content layer is modified to accommodate the new domain, namely nineteenth-century German documentary sources on the Illuminati order, with FactGrid as the linking target. Demonstrated success in this context substantiates the transferability claim articulated in RQ3 and affirms COMP-PROV as a transferable design commitment (C4) rather than a Vasari-specific artifact. Extension to multiple-case replication across additional source types is identified as a necessary condition for further strengthening this claim, as discussed in Section 12.3.

10.2 The Knigge case study: applying ObliquER to the Kloß' *Annalen*

This section operationalizes the transfer. Section 10.2.1 describes the primary source, the principled excerpt, and the wider annotated corpus and schema; Section 10.2.2 reports the recognition results.

10.2.1 Corpus and source selection

The primary source is Georg Burckhard Kloß' *Annalen der Loge zur Einigkeit der Englischen Provincial-Loge, so wie der Provincial- und Directorial-Loge des eclectischen Bundes zu Frankfurt am Main: 1742–1811* (Frankfurt am Main: Horstmann, 1842; FactGrid Q11399). It is a retrospective chronicle of lodge activities compiled from archival records, membership registers,

and correspondence, written in nineteenth-century German and structured chronologically. As a genre, such lodge annals are documentary in origin – the modern reconstruction of the Illuminati order rests precisely on this kind of internal archival record (Agethen 1984) – yet, like any source, they are not free of interpretation: the *Annalen* are a retrospective, selective compilation whose framing must itself be subjected to source criticism (Droysen 1868, Bernheim 1908, Fickers 2020). The very secrecy that the lodge records document is also what lent the order to conspiratorial reading. Koselleck (1959) traces how the structural opacity of the Enlightenment secret societies generated a politics of suspicion, and the Illuminati in particular became the anchor of a durable conspiracy-theoretical tradition (Oberhauser 2013). The digitization is accessible via the Bavarian State Library, digitized via the Münchener Digitalisierungszentrum Kloß (1842) and is cited by FactGrid as a source for multiple lodge-membership and visiting claims, including the claim that Adolph Freiherr von Knigge (Q595) visited the Lodge Zur Einigkeit / L'Union (Q11161) between 1778 and 1781.

Corpus selection and ground truth. Similar to the GT construction and sampling process presented in Section 7.4.4 for Vasari's work, we built the GT from numbered paragraphs, anchored to a core excerpt (which we analyzed in detail), and then expanded into a controlled stress test. The core excerpt is § 174 (*Der Illuminatenorden zu Frankfurt*), § 175 (*Gogels letzte Arbeiten*), and § 176 (*Schilderung von Gogels Character*), pp. 135–137 of the 1842 edition, the primary extraction target from which the worked KG example of Section 10.2.3 is built. Given the lack of a modern critical edition of the *Annalen*, we selected these examples from existing triples in FactGrid. The choice of the core window was further guided by the author's prior work on the NFDI4Memory / MemO KG (Ondraszek et al. 2025a,b), which provided the context for this transferability study (cf. Section 10.1) and established familiarity with both the FactGrid Illuminati records and the passages of the *Annalen* that FactGrid documents most densely. This provides the silver standard against which the extraction can be evaluated, thereby anchoring the core window on the most densely documented passages. The selected passages form a complete narrative episode (the Illuminati Order's entry into the Frankfurt lodge network, Knigge's recruitment by Costanzo di Costanzo, his dated first and last visits of 3 July 1778 and 16 September 1781, and the cell's later trajectory under Peter Gogel). Additionally, we found § 174 to be of particular suitability because it contains all five ablated mention types, which we will elaborate on in Section 10.2.1: explicit named mentions (*Br. Adolph v. Knigge*), coded aliases (*Philo*, *Diomedes*), coreferential pronouns (*Dieser*), implicit organization references (*seiner Ordnung*), and OOKB persons (*Br. Herophilus*).

Almost all of the mentions and entities to be inferred from these passages are existing FactGrid claims, resulting in a structured GT for EL that is available without additional manual annotation. For example, this is evident in FactGrid's record of Knigge's role as *Visitor of the Lodge* at Q11161

between 1778 and 1781 (Property P91; role Q172997), citing page 135 of the *Annalen* as one of two sources. In the future, the EL output of ObliquER can therefore be evaluated against a pre-existing structured claim.[3]

We embedded this core set of passages, along with corresponding page images, together with five further thematic windows. These were processed with Tesseract OCR (Smith 2007) using a Fraktur-trained model (`deu_latf+Fraktur`[4]) and manually corrected against the facsimile (Fraktur ligatures and long-*s* resolved, standard abbreviations expanded except referentially load-bearing titulars), yielding a 78-entry paragraph pool. Removing section headings and paragraphs shorter than 80 characters, uniformly applied across all six windows, resulting in 64 annotated paragraphs (cf. Table 10.1), sampled from six thematic windows spanning pp. 134–315:

A. **§§ 173–176 (pp. 134–137) – core Illuminati window.** Contains the core excerpt of Section 10.2.1 together with § 173 (which supplies the few-shot example); it is the only window annotated for all five ObliquER mention categories.

B. **§ 181 (p. 142) – independence declaration.** Administrative register, dense organizational and locational inventory, no coded aliases – a contrast to the conspiratorial register of window A.

C. **§ 196 (pp. 158–159) – Freymaurer-Clubb.** Knigge appears by real name rather than alias, a contrastive case for the alias-versus-explicit distinction.

D. **§§ 199–201 (pp. 164–167) – Ditfurth episode.** The second coded-mention test bed (Ditfurth / *Minos*), with the highest alias density outside the core window and multiple simultaneous alias chains.

E. **§§ 293–294 (pp. 276–277) – Gotha circular (1791).** Administrative epistolary prose at temporal and thematic distance from the core narrative.

F. **§§ 329–331 (pp. 314–315) – *Vorfälle in der Loge* (1803–1804).** A highly varied entity landscape (individual admissions from Speyer, Breslau, Nuremberg, and Chemnitz) sixty years after the Illuminati window, including a documented anti-Jewish admission refusal (§ 330).

Distribution of challenge categories. With each window, we aim to isolate a distinct challenge for the extraction. For instance, this concerns coded aliases, which are confined entirely to 26 in core window A and 13 in window D. Furthermore, 12 of the 14 OOKB mentions are

[3] Claim in FactGrid: `https://tinyurl.com/382ryyyn`
[4] See: `https://ocr-d.de/en/models.html`.

concentrated in window F, a section focused on the region of marginal members that are absent from FactGrid. FactGrid-tagged mentions co-localize with the coded-alias windows, with 25 in A and 17 in D. We found this pattern to reflect the documentary character of the source text, where entities cataloged by FactGrid correspond to those documented in the Annalen, given that FactGrid already gives these texts as sources for their claims (cf. Section 10.2.1). Consequently, the corpus tests the extraction architecture across a spectrum that ranges from silver-standard-supported linking, as in windows A and D, where FactGrid provides most identifiers, to genuinely independent OOKB identification in window F.

Annotation schema. We extend the ObliquER annotation schema (see Section 7.4.6) with new entity types, namely coded alias, which denotes an Illuminati pseudonym whose referent is present in the knowledge base but whose surface form is not accessible to standard named entity recognition (for example, Philo refers to Knigge, Minos to Ditfurth); coded location, which designates a code name for a place (for example, Edessa refers to Frankfurt am Main); and explicit organization, which identifies a lodge, order, or sub-body named directly in the text and replaces the Vasari-specific explicit artwork mention. Additionally, the schema now features types for explicit location and explicit event; we modify the inclusion of explicit person mention, retain coreferent, and add implicit organization.

Annotation process. We annotated the GT in one run[5] against the adapted schema, reading each of the 64 paragraphs in source order and resolving identifiers by direct lookup against FactGrid. Our current pilot links to FactGrid only, since it contains all the specialized entities, but it foresees dual linking to Wikidata. This has a direct impact on transparency, since for most entities, the inference of an entity is not an independent identification act but the *recovery* of an assignment that a FactGrid editor has already made against the same passage. For example, Q595 cites § 174 (p. 135), and the lodge (Q11161), Costanzo (Q217), the Illuminati order (Q10677), and the author (Q1459) are similarly cited from the same text or from sources Kloß compiles. Of the 48 mentions carrying a FactGrid ID (Table 10.1), the majority are of this kind, so our contribution is closer to documentation than to independent identification. The 14 OOKB mentions, for which no FactGrid item exists, are the only category whose identifier assignment is a genuinely independent decision. Coded-alias resolutions were cross-checked against Bois (1993). As a Yin-style theoretical replication, the pilot's evidential weight comes from controlled variation against the Vasari corpus on the dimensions the prompt content layer must absorb (language, period, register, entity types, KB), not from a saturating annotation study; a multi-coder pass with formal inter-annotator agreement on the same corpus is identified as future work (Section 12.3).

[5] The GT was annotated, with FactGrid as a reference, by the author of this work.

Adapted prompts. To test the transferability of ObliquER's architecture, we change the prompts to use the identical structural template as those for Viewsari (Section 7.4.8.2), with domain-specific content added. As for role prompting, we instruct it to function as an expert in eighteenth-century German sources on secret societies; the artwork-focused entity definition is replaced with the schema mentioned above. We extend the task to cover FactGrid-targeted OOKB detection, coded-alias and coded-location decoding recorded in a label field, and parenthetical alias pairs ("Br. Schmerber (Agathoeles)") annotated as two mentions – one explicit, one coded – sharing one entity identifier; and the output schema adds a FactGrid-identifier field, a label field, and an in-text span-annotation field. A single extended few-shot example from § 173 demonstrates the five hardest patterns at once, namely the *Diomedes*/*Philo* alias pair decoded to Q217/Q595, a parenthetical pair, the OOKB *Br. Herophilus*, the *Unionsloge* → Q11161 organization mention, and cross-paragraph coreference via the preceding paragraph's output. Our ObliquER-Illuminati pilot runs this single few-shot configuration; the ablation variants compared to Viewsari (Section 9.4.6) are not re-instantiated because the previous evaluation (Section 9.4.6) has already isolated the contribution of demonstrations from that of the ontology-typed content layer.

Corpus statistics. Table 10.1 summarizes the annotated corpus.

Table 10.1: Summary statistics for the Kloß *Annalen* annotated corpus.

Metric	Value
Total paragraphs sampled	64
Total annotated mentions	237
Explicit person mentions	99
Explicit organization	52
Explicit location	39
Coded alias	39
Implicit organization	5
Coded location	2
Coreferential	1
OOKB entities	14
Mentions with FactGrid ID	48

The 39 coded aliases across 64 paragraphs confirm the high alias density of the Kloß *Annalen*, and the 14 OOKB entities – persons documented in the text but absent from FactGrid – are the discovery contribution of the pipeline for this corpus, a parallel to the OOKB artwork entities in

the Vasari study. The raw paragraph pool, the structured 64-paragraph sample, and the annotated GT are available in the project repository.

The source genre exhibits a higher density of explicit entity mentions, a pattern that stands in marked contrast to Vasari's work. The documentary character of lodge chronicles, which are fundamentally structured around the enumeration of named individuals, lodges, locations, and dates, leads to a predominance of explicit references. This configuration reverses the Vasari scenario. Rather than seeking a balance between explicit and implicit mentions, the focus here is on achieving comprehensive coverage with quantifiable density. Windows A and D are selected to ensure that the coded-alias category, the structural counterpart to Vasari's implicit mentions, appears frequently enough to permit systematic evaluation. Window F is included to enable the identification of genuinely OOKB entities. Explicit categories are incorporated at the frequency determined by the inherent properties of the genre.

10.2.2 ObliquER recognition results on the Kloß corpus

We ran the adapted few-shot prompt template over the 64-paragraph Kloß GT (Table 10.1) under the two inference models of the Viewsari cross-model evaluation (Section 9.4.5), `openai/gpt-oss-120b` and `google/gemma-4-31b-it`. We performed only recognition with ObliquER, scoring span-level identification and mention typing under the SemEval 2013-9.1 criteria (Segura-Bedmar et al. 2013). We excluded an EL step from this run and reserved it for future work (Section 12.3) because the existing linker focuses on Wikidata rather than FactGrid.

Table 10.2 reports performance over the 236 scored gold mentions.[6]

We observe that the coded alias (the Kloß structural analog of the implicit artwork mention, in which an Illuminati pseudonym denotes a referent whose surface form is opaque to standard recognition) is the highest-scoring populated type under both models (Strict F1 $= 0.828$ on `gpt-oss-120b`, 0.878 on `gemma-4-31b-it`). The non-transparent-reference capability that distinguishes ObliquER on Vasari (Section 11.3) therefore transfers to the conspiratorial-pseudonym mechanism of the *Annalen*, which is the central recognition claim the pilot is designed to test.

Moreover, the cross-model direction observed on Viewsari replicates what we have established in Section 9.4.5. `gemma-4-31b-it` edges out `gpt-oss-120b` on the overall Strict F1 (0.683 vs. 0.600) and Type F1 (0.733 vs. 0.685) under the held-constant prompt, consistent with the cross-model envelope of Section 9.4.5. The reading of the coded-alias figure follows Section 11.4,

[6] The 64-paragraph GT carries 237 annotated mentions (Table 10.1); the single coreferent mention is not scored as its own type in the recognition evaluation, leaving 236 gold mentions across the seven recognized types.

Table 10.2: ObliquER recognition on the 64-paragraph Kloß *Annalen* GT (236 scored gold mentions), under the single adapted few-shot prompt held constant across two inference models. F1 is reported under the SemEval 2013-9.1 span-level criteria (*Strict* = exact span, exact type; *Partial* = half credit for overlapping spans; *Type* = correct type with overlapping span).

		`gpt-oss-120b` F1			`gemma-4-31b-it` F1		
Mention type	n_{gold}	**Strict**	**Partial**	**Type**	**Strict**	**Partial**	**Type**
Overall	236	**0.600**	**0.739**	**0.685**	**0.683**	**0.763**	**0.733**
Coded alias	39	0.828	0.828	0.828	0.878	0.878	0.878
Coded location	2	0.571	0.571	0.571	0.667	0.667	0.667
Explicit person mention	99	0.704	0.833	0.833	0.843	0.898	0.898
Explicit location	39	0.674	0.719	0.719	0.716	0.737	0.737
Explicit organization	52	0.516	0.661	0.661	0.500	0.598	0.598
Implicit organization	5	0.000	0.000	0.000	0.167	0.167	0.167

showing that the coded alias is an easier recognition target than the Vasari implicit mention, a single opaque token rather than a descriptive paraphrase, and three of the recurring aliases are demonstrated in the prompt itself, so the pilot supports the transfer of the non-transparent-reference capability *as a class* rather than at equal difficulty. The *generic collective mention* type has no gold instances, so its predictions (26 under `gpt-oss-120b`, 25 under `gemma-4-31b-it`) form a pure false-positive class with F1 $= 0$.

The error profile is concentrated in over-prediction. *Implicit organization* (5 gold) collapses to F1 ≈ 0 with 33 predictions under `gpt-oss-120b`; and *explicit organization* is the weakest well-populated type (Strict F1 ≈ 0.50, 72 predictions under `gpt-oss-120b` and 112 under `gemma-4-31b-it` against 52 gold). The overall figures (Strict F1 0.600 / 0.683) sit above the Vasari artwork-recognition figures (Strict F1 0.334, Table 9.7) because the Kloß corpus is dominated by explicitly named persons, locations, and organizations rather than by the descriptive implicit paraphrases that make the Vasari artwork surface hard; the cross-corpus comparison is therefore directional rather than like-for-like, and the result to read is the within-corpus standing of the coded-alias type.

10.2.3 From recognition to a knowledge graph

The transfer results of this chapter, that ObliquER's recognition surface carries over to the *Annalen* (Section 10.2.2) and that the Viewsari ontology core absorbs the new entity and mention types (Section 10.3.2), combine into the chapter's central claim that the two components together constitute a working pipeline whose output can serialize to a KG on a corpus that shares none

of the characteristics of Vasari's work. The goal of the pilot is not to deploy a full Kloß KG as a parallel to FactGrid, but to provide an example under the Viewsari ontology (Figure 10.1), serving the same illustrative purpose as the visualization in Figure 5.7. This graph shares the same class hierarchy, the same dual `oa:Annotation` / `prov:Entity` typing of every mention, the same `doco:TextChunk` / `oa:TextPositionSelector` target structure, and the same provenance attribution, with only a change to the mention type taxonomy. In total, the graph consists of 5,152 triples over the 64-paragraph corpus, containing the 237 typed mention annotations across the three-layer structure. The schema extensions (`kx:coded_alias`, `kx:coded_location`, `kx:explicit_organization`) are declared as subclasses of the existing Viewsari mention classes.

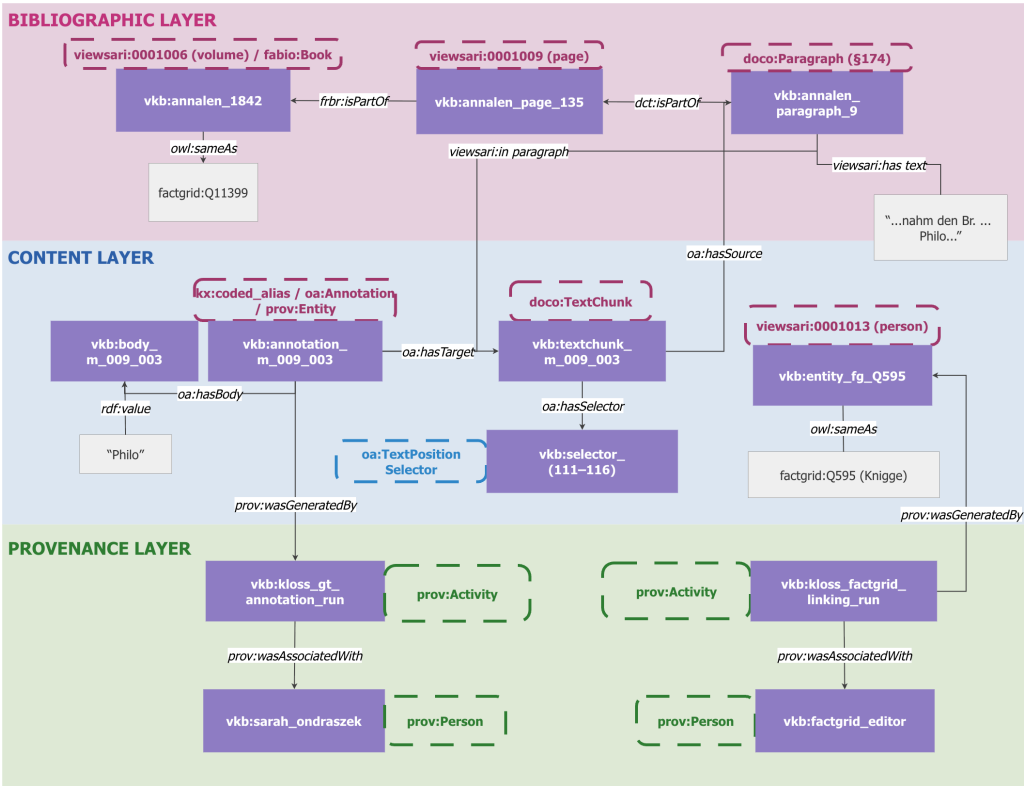


Figure 10.1: Worked example of the full three-layer Viewsari representation of a single Kloß *Annalen* extraction, taken from the mock KG. The coded alias *Philo* in § 174 is modeled across the bibliographic layer (the *Annalen* 1842 volume, `owl:sameAs` FactGrid Q11399, with its page and paragraph), the content layer (the `kx:coded_alias` annotation, its text chunk and position selector, and the resolved entity `owl:sameAs` FactGrid Q595), and the provenance layer (the typed extraction and identifier-recovery activities attributed to their respective agents). The structure is identical to the Vasari instantiation; only the domain content differs.

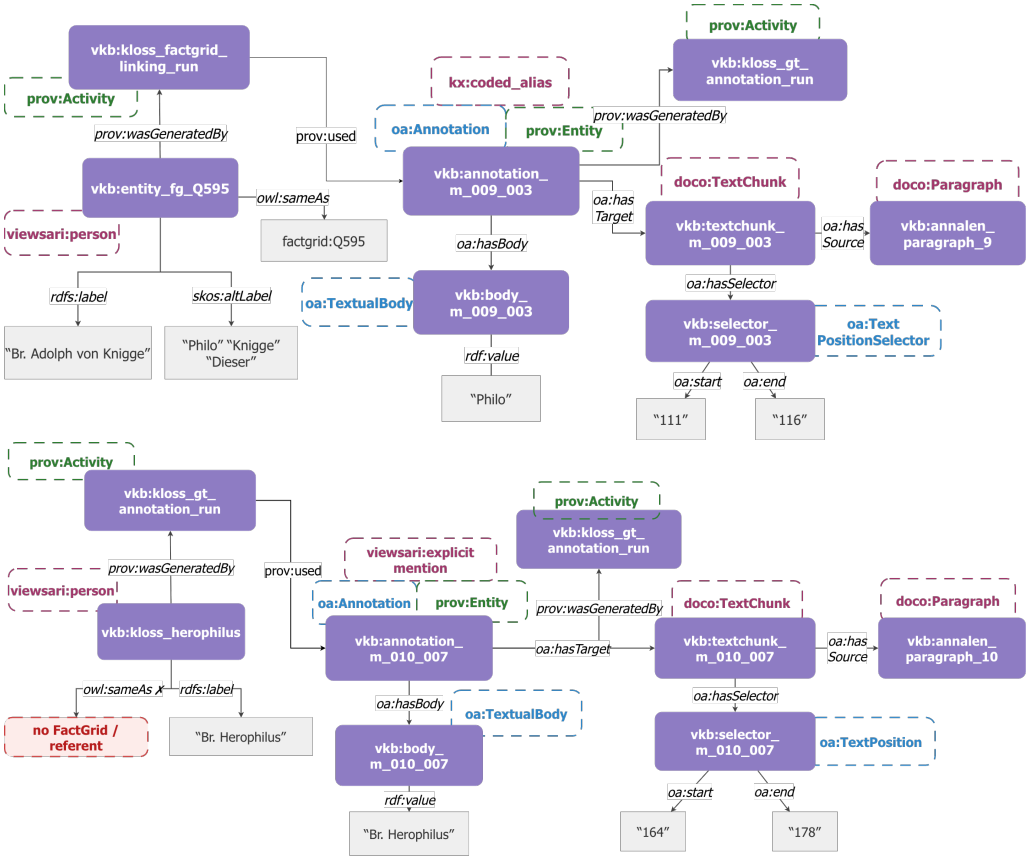


Figure 10.2: The resolved/OOKB bifurcation on two Kloß entity nodes (mock graph). *Top:* the coded alias *Philo* resolves to a person entity linked `owl:sameAs` FactGrid Q595 (Knigge), with its surface forms *Philo*, *Knigge*, and *Dieser* carried as `skos:altLabel`. *Bottom:* the explicitly named *Br. Herophilus* has no FactGrid or Wikidata referent and is retained as an OOKB entity with a stable internal IRI – the pipeline's discovery contribution for this corpus. Both nodes carry the same dual `oa:Annotation`/`prov:Entity` mention typing and full provenance.

The example renders the dual-provenance scenario from Figure 5.7 as concrete triples. In this instance, the Knigge and Lodge co-occurrence on Section 174 serves as the shared `oa:hasBody` for two annotations with a common `oa:hasTarget`, taken from the mentions of these entities in the same texts. One annotation is attributed to a human FactGrid editor (`prov:Person`), and the other to an ObliquER run (`prov:SoftwareAgent`); these differ only in `prov:wasGeneratedBy` (see Figure 10.6). The FactGrid entity linking step, described in Section 12.3, is the only part of the process that has not yet been fully evaluated.

10.3 Transferable design patterns – Viewsari for other humanities knowledge graphs

This section presents a comparative study of ODPs in four digital humanities projects: ICON (Sartini et al. 2023), Polifonia (de Berardinis et al. 2023), ODEUROPA (Lisena et al. 2022), and Viewsari (Ondraszek et al. 2024c). These projects were chosen for their range of domains, including artistic interpretation, music heritage, olfactory heritage, and biographical-artistic literature. Selection also considered the maturity of their ontologies and how they address provenance and interpretive uncertainty. For each project, we reviewed the published ontology documentation and competency questions, identified the ODPs they use or instantiate, and examined their design methods. ICON, Polifonia, and Viewsari all use the XD methodology (Blomqvist et al. 2016).

Author contributions. The comparative study was conducted jointly with Bruno Sartini, Marieke van Erp, Andrea Poltronieri, Pasquale Lisena, and Harald Sack. The author of this dissertation contributed the Viewsari case-study analysis, the ANN pattern formalization, the comparative table, and the transferability argument; the ICON analysis was led by Sartini, the Polifonia analysis by Poltronieri and Lisena, and the ODEUROPA analysis by van Erp. The paper was written collaboratively, with the author contributing the introduction, the Viewsari section, and the discussion.

10.3.1 From case studies to comparative patterns

Our finding (Table 10.3) is that the ten identified pattern categories are largely project-specific. Only PROV is shared across most case studies, confirming that DH ontologies remain substantially domain-tailored despite drawing on a common core of reference ontologies (such as CIDOC CRM (Doerr 2003) or the IFLA family (IFLA Study Group on the Functional Requirements for Bibliographic Records 2009)). This is not a failure of the field; it reflects the genuine

heterogeneity of humanities data and the inadequacy of any single overarching model to express the full epistemic register of each domain. However, we observed, in particular, that there are still commonalities among them. The annotation pattern (ANN), which we abstracted from the Viewsari representation of co-occurrences as first-class annotated objects with additional metadata, shows broader implicit applicability. For example, an interpretive act in ICON could be modeled as annotated statements with provenance chains, without domain-specific specialization at the pattern level. In relation to the provenance chains, our survey exposed a gap that none of the four projects have previously addressed, given in the form of no existing DH ontology pattern for the provenance of *computationally generated* interpretations. This refers to cases where a machine agent, such as an LLM, rather than a human scholar, makes the annotation.

Our collaborative work in 2024c generalizes the ANN pattern beyond its initial implementation in Viewsari. In this pattern, an annotation attaches to a *statement* and records five preliminary dimensions, namely provenance, purpose, place, time, and agency. This structure ensures that any assertion can be traced to the specific conditions under which it was created. When considered alongside the four case studies, this abstraction clarifies the connection to COMP-PROV. Specifically, COMP-PROV is obtained by integrating the aforementioned dimensions into a dedicated specialization. While the ANN abstraction (Figure 10.4, left) is (implicitly) encoded into three of our ontologies, the PROV pattern is the only one common to all four case studies. We identified this as a unified PROV-O provenance layer for computationally assisted extraction (Figure 10.4, right). ODEUROPA provides the most relevant precedent because the EOKG employs PROV-O to associate each smell, smell-emission, and olfactory-experience with the textual or visual fragment from which it was extracted (Lisena et al. 2022). Returning to the ANN pattern, we can hold that, essentially, this approach instantiates the provenance and place/time dimensions of the abstract pattern for both human and rule-based extractors, as required by approaches like ODEUROPA, as shown in Figure 10.3.

We compose COMP-PROV to address this limitation by binding the ANN pattern's agency to a parameterized software agent that records the run identifier, prompt template, and model configuration. Through the explicit modeling of agency, COMP-PROV completes the ANN pattern. As a result, COMP-PROV integrates with ODEUROPA's existing PROV-O layer (Figure 10.7) rather than competing with it.

However, ODEUROPA, along with ICON's interpretation acts and Polifonia's attributions, makes no distinction between the textual mention produced by a software agent and the resulting referent. Figure 10.3 shows that ODEUROPA collapses the textual mention (the linguistic sign in the source text) and the olfactory phenomenon (the smell itself) into a single node. This is PIEA operative in a deployed KG; referent and mention are collapsed into unison, and the mention does not survive as a first-class object that resolution could leave open. If, for example, we have a text

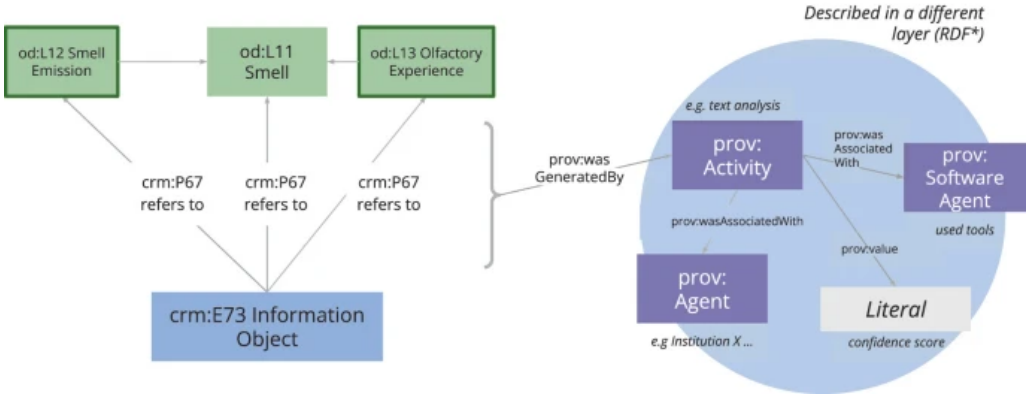


Figure 10.3: The provenance of information represented in the ODEUROPA data model, illustrating how olfactory concepts (`od:L12 Smell Emission`, `od:L11 Smell`, `od:L13 Olfactory Experience`) relate to `crm:E73 Information Object` via `crm:P67 refers to`, and how provenance is captured in a separate layer linking the information object to `prov:Activity`, `prov:Agent`, `prov:SoftwareAgent`, and associated literals (e.g., confidence scores) through PROV-O properties. Figure as shown taken from Lisena et al. (2022)

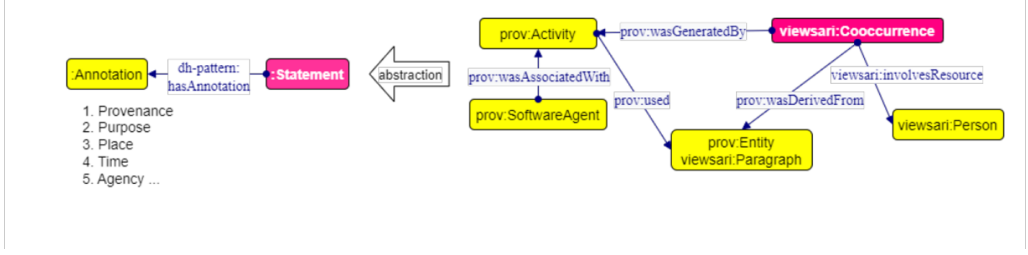


Figure 10.4: The ANN pattern and its extension toward COMP-PROV, after Ondraszek et al. (2024c). Left: the abstract Annotation pattern – an annotation attached to a statement, recording provenance, purpose, place, time, and *agency*. Right: its Viewsari instantiation, in which the agency and provenance dimensions are realized with a PROV-O layer and a typed `prov:SoftwareAgent` – the merge that COMP-PROV formalizes for computationally-assisted extraction. Because the right-hand structure only adds a software agent to a PROV-O provenance chain, it applies directly to any DH KG already using PROV-O (e.g. ODEUROPA) and to the Kloß transfer (Figure 10.6).

saying "She noticed that the rose that he gifted her smelled sweet," ODEUROPA would link the smell directly to the information object (the text), but currently does not carve out an intermediate layer for "the token in position X referring to this smell concept". Since this is exactly the kind of distinction our PIEA critique targets (the model assumes that the entity in the KG is straightforwardly what the text is about), we overcome this issue in the COMP-PROV pattern by retaining the annotation-output mention layer.

Table 10.3: Updated comparative pattern analysis across four DH case studies, extended with the COMP-PROV pattern introduced in this dissertation. ✓ = pattern explicitly employed; (✓) = pattern implicitly instantiated or applicable; × = pattern absent. Abbreviations: Situation (*SIT*), Description (*DESC*), Visual Semiotics (*VIS-SEM*), Interpretation (*INT*), Information Realization (*IR*), Provenance (PROV), Person (*PER*), Place (*PL*), Annotation (ANN), Analog vs. Digital (*AN-DIG*), Computational Provenance (COMP-PROV).

	SIT	*DESC*	*VIS-SEM*	*INT*	*IR*	*PROV*	*PER*	*PL*	*ANN*	*AN-DIG*	*COMP-PROV*
ICON	✓	✓	✓	✓	×	(✓)	✓	✓	(✓)	✓	×
Polifonia	×	×	×	×	✓	✓	×	×	(✓)	×	×
ODEUROPA	×	×	×	×	×	✓	×	×	×	×	×
Viewsari	×	×	×	×	×	✓	✓	✓	✓	✓	✓

10.3.2 The Computational Provenance pattern

COMP-PROV provides a reusable pattern for contexts where LLM-based extraction populates a humanities KG. As established in Section 9.2, among the standard schemas for CH applications, neither CIDOC CRM, CRMdig (Doerr et al. 2016), nor CRMinf (Doerr et al. 2023) provides adequate vocabulary to record the epistemic character of a computationally generated annotation. This concerns not only agency and further parametric configurations but also input spans and certainty values. Crucially, the pattern holds the textual or visual *mention* – the annotation output the activity generates from its input span – distinct from the *referent* it resolves to; this intermediate layer, whose absence is exactly what licenses PIEA, is what the pattern interposes between the source and the resulting entities. Because that mention node carries the activity's provenance (agent, run, prompt template, model configuration), the interpretive step from source to the resolved entity is retained as an auditable claim rather than compressed into a bare assertion that the knowledge-base entity is simply what the source is about. It is an abstract formulation of the provenance trail implemented in Viewsari, aligned with observations from the other three

ontologies and the FactGrid transfer pilot, and made transferable by removing all Viewsari-specific namespace bindings and expressing the structure in generic terms.

Essentially, this pattern is a specialization of Viewsari's ANN pattern, and the finding that PROV is the one design commitment shared across the DH case studies was *merged*, over the course of Viewsari ontology development, into a single PROV-O specialization for computationally assisted extraction. The Kloß pilot provides one of the areas where the merged pattern holds, alongside ODEUROPA.

Abstract pattern definition and transfer to other examples. Figure 10.5 shows the abstract COMP-PROV pattern.

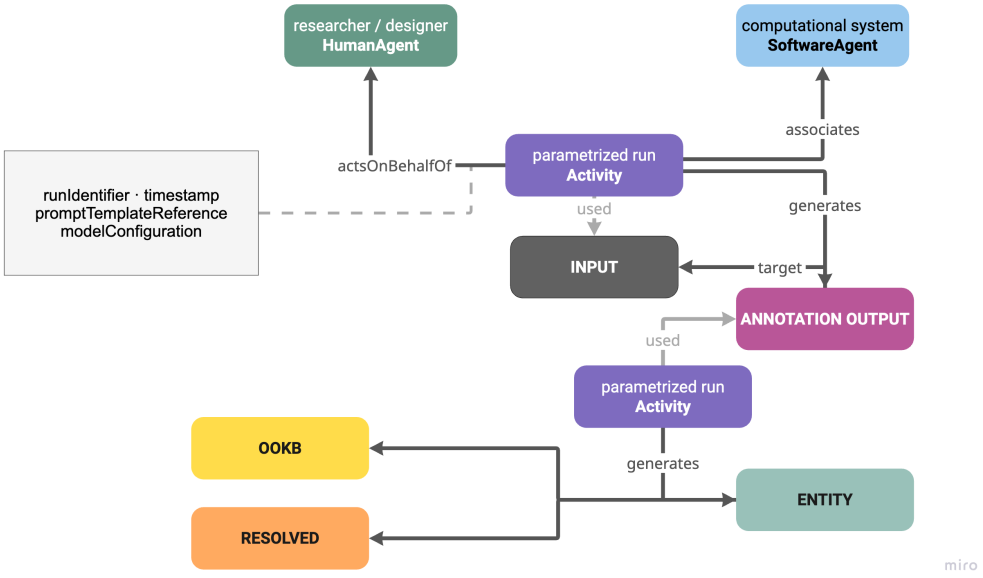


Figure 10.5: The abstract COMP-PROV pattern. A `HumanAgent` (researcher/designer) `actsOnBehalfOf` a parametrized `ExtractionActivity` that a `SoftwareAgent` `associates` with; the activity `uses` an `Input` chunk and `generates` an `AnnotationOutput` – the mention layer – whose entity (`wasDerivedFrom` the output) resolves either to a KB referent or to an OOKB entity carrying a stable internal identifier. Holding this mention distinct from the referent is what keeps the resolution an attributable claim rather than a bare assertion. The attribute box records the run identifier, timestamp, prompt-template reference, and model configuration.

Figure 10.6 shows the Viewsari-based instantiation of COMP-PROV on the Kloß *Annalen*, in which the abstract roles are bound to concrete ontology terms. This instantiation foregrounds

the mention/referent separation. The resolved entity (a FactGrid or Wikidata referent) is never asserted directly of the text but reached only through an `oa:Annotation` *mention* anchored to a source span, and it is onto that shared mention layer that the provenance attaches – here split into two parallel chains, one human-made and one computational.

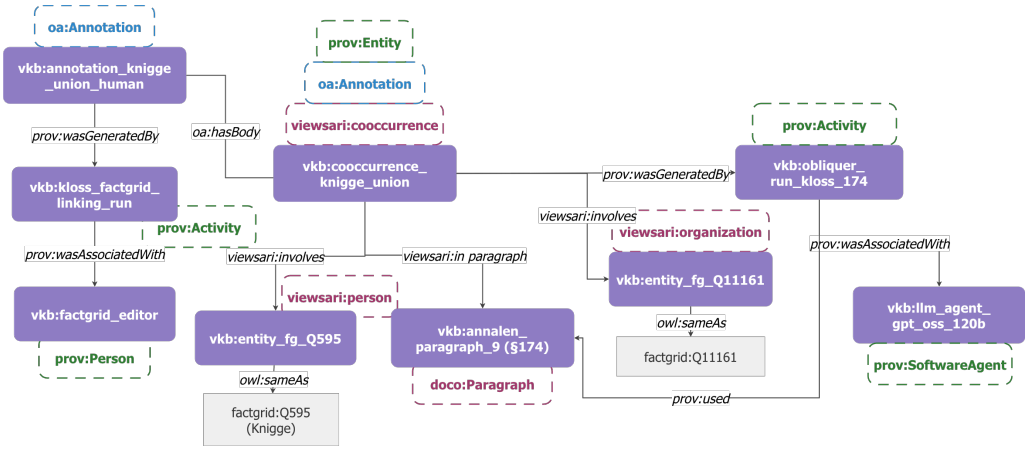


Figure 10.6: Mock-KG inspired rendering of the dual-provenance co-occurrence from the persons and organizations extracted with ObliquER (cf. the schematic in Figure 5.7). The co-occurrence of Knigge (`owl:sameAs` FactGrid Q595) and the Lodge Zur Einigkeit / L'Union (Q11161), asserted on § 174, is the shared `oa:hasBody` of two annotations with a common `oa:hasTarget`. A human FactGrid-editor annotation (left, attributed via `prov:wasAssociatedWith` to a `prov:Person`) and an ObliquER annotation (right, attributed to a `prov:SoftwareAgent` and recording the prompt as a used `prov:Entity`). The two layers compose without competing for the same slot.

For the same reference, the pipeline automatically adds extracted mentions (cf. Section 10.2.2). As shown in Figure 10.2, ObliquER creates a parallel annotation for the same text chunk expressing the co-occurrence of Knigge and the lodge, with its provenance activity linked to a `prov:SoftwareAgent` and its prompt-template hash and model version recorded on the activity node. Both annotations share the same `oa:hasTarget` and linked entities and differ only in `prov:wasAssociatedWith` – one resting on scholarly editorial authority, the other on a documented LLM extraction run.

The same construction applies to ICON (Sartini et al. 2023), the one comparison ontology whose explicit purpose is to model *interpretation*. ICON formalizes the reading of an artwork across Panofsky's three recognition levels and, in its native form, attributes each recognition act to a *human* interpreter. Once the interpreter is a vision–language model rather than a scholar, in efforts such as the Prado project (Barcelona Supercomputing Center 2023), the direct probing for art-historical competence (Strafforello et al. 2025), and IICONGRAPH's anticipated LLM-based question answering over iconological graphs (Sartini 2024), becomes

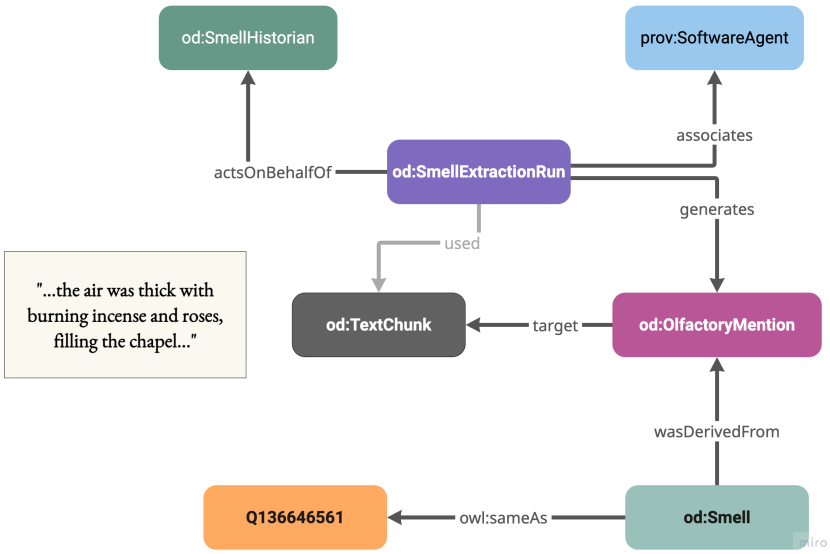


Figure 10.7: Hypothetical instantiation of the COMP-PROV pattern in the ODEUROPA / *EOKG* context. The abstract roles bind to ODEUROPA-specific (fictional) classes, and the textual mention from which the concrete smell entity is later derived is marked as *AnnotationOutput*, anchored to the passage. `od:Smell` is then the resolved entity.

exactly the gap COMP-PROV closes. Binding ICON's `icon:IconographicalRecognition` to the `ExtractionActivity` role, with a `prov:SoftwareAgent associated` and the art historian who fixes the run `actingOnBehalfOf` it, the recognized subject is retained as an `AnnotationOutput` mention, so a separate instance that resolves to an ICONCLASS or Wikidata referent. The interpretive act is thereby preserved with its provenance detail. A computationally generated reading remains legible as a claim attributable to a specific model under specific parameters and is auditable and revisable, rather than being silently merged into the artwork's asserted iconography.

10.3.3 The co-occurrence content pattern

Where COMP-PROV records how an annotation was produced, the second pattern contributed here concerns what a recurring class of annotations asserts. We briefly reintroduce the co-occurrence pattern, which is itself a member of the ANN family (an annotation on a statement about a source span) and can therefore carry a COMP-PROV provenance chain like any other annotation. Essentially, it describes an entity-entity relation from textual observation. Polifonia's MEETUPs make the composition concrete, extracting musician meetups from text (Tirado et al.

2023). However, MEETUPs inherits the Participation pattern[7] and types it as an event with participants, collapsing the textual mention into a real-world referent. A joint mention of two musicians in a source is taken as evidence of a meeting between them. The pattern set out below is the corrective, holding the relation on the side of *what the text records* rather than *what happened in the world*.

Co-occurrence, accordingly, addresses a recurring modeling problem in DH and network analysis, namely how to represent and trace *textual or computational observations* that two or more entities have been encountered together, while preserving the evidence on which the observation rests and without committing the ontology to any worldly relation between the participants. The distinction matters because a co-occurrence is not an event, and two textual mentions named together in a paragraph are observations about *texts and records*, not assertions about meetings, collaborations, or shared activity in the world.

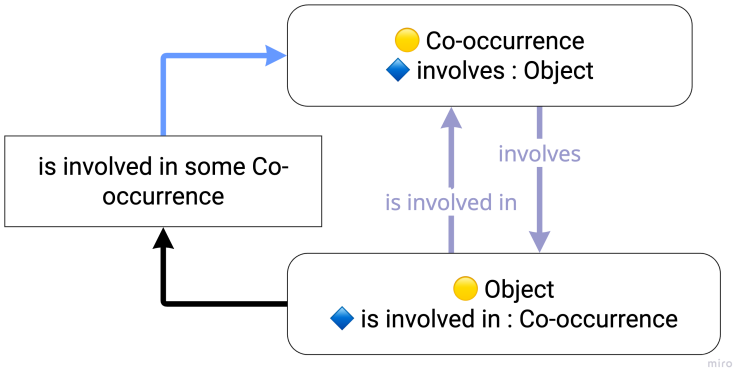


Figure 10.8: ODP for co-occurrence modeling. A `Co-occurrence` class (yellow circle) holds an `involves` object property ranging over `Object` instances (blue diamond), while each `Object` exposes an inverse `isParticipantIn` property linking back to an `Event`. The existential restriction *is involved in some Co-occurrence* (left, black arrow) captures the OWL 2 necessary condition that every participating object must be linked to at least one co-occurrence instance.

The pattern proposes that a co-occurrence be modeled as a first-class entity in the provenance-bearing sense. It is an artifact of an extraction or observation process. Figure 10.8 illustrates the design of the co-occurrence pattern. We offer this pattern as an alternative to the Participation pattern, which we deprecated from the Viewsari ontology. Participation reifies events of joint action; Co-occurrence reifies records of joint mention.

[7] We introduced this Pattern in Section 7.3.1 and incorporated it into the first version of the Viewsari ontology. The pattern is event-centric and does not treat textual co-occurrences or relation extraction results.

11 Discussion

With PIEA, we have explicitly expressed what conventional KG construction handled implicitly. Mention and referent are collapsed, filling KGs with seemingly settled facts rather than context-dependent claims. The irony, given the case study this dissertation chose, is that implicitly conveyed information is precisely the focus of our work. The references that matter most in *The Lives*, the artworks Vasari names only by description, are implicit, and the extraction process to obtain them from the text with generative models is interpretive in a newer sense. The work gathered in the preceding chapters identifies PIEA and its rejection in the Viewsari ontology, collects data through ObliquER and its mention typology, and documents the construction process through a provenance model that records the conditions under which each claim was made.

In this chapter, we put together the puzzle pieces from those chapters against the problem observed in Chapter 1 to form the full picture. The problem statement gathered that under generative extraction, where a model constructs interpretations rather than detecting facts, *extraction without documentation is epistemology without accountability*. Correspondingly, in Chapter 9, we report our evaluative findings together with brief interpretive cues (our puzzle pieces), which we here fit together to discuss their interplay against the central claim of the dissertation to answer the three RQs of Section 1.3.

Knowledge extracted from interpretive texts is not discovered but constructed; therefore, knowledge graphs must model and make visible the construction process itself.

In the following sections, we frame the overarching argument in Section 11.1; Section 11.2–11.4 address each RQ in turn, drawing on evidence from the results (Chapter 8), evaluation (Chapter 9), and transferability study (Chapter 10).

Section 11.5 consolidates the comparative reading of CRMinf (Doerr et al. 2023) and Viewsari, and Section 11.6 addresses the recurring objection that agentic LLMs render the Viewsari modeling effort obsolete. Throughout, each answer is held to the criterion of answerability stated for its question in Section 1.3. Section 11.7 opens up a reading of an unreliability objection of creating KGs from interpretive, uncertain texts, and Section 11.8 reflects on the methodological lessons learned.

11.1 When interpretation becomes an epistemic act of large language models

Our central claim, that knowledge from interpretive historical texts is not discovered but constructed, and KGs must therefore model and make the construction process itself visible, is a guiding light from Chapter 1 on. Along four dimensions, we add to it with the results from our pipeline, their evaluation, and their contextualization under the lens of this claim.

We begin this part with the observation that implicit reference and absence from contemporary KBs track each other structurally. The entities Vasari assumes his readers will recognize without naming are precisely the entities for which no canonical identifier was ever established. A pipeline restricted to explicit mentions would lose not just a fraction of the artwork population but the specific subset most tightly bound to the canonical narratives *The Lives* itself helped establish. This is not a data quality problem; it is a canonization signature preserved in the KG (Graciotti et al. 2025).

The distribution of the GT reveals this mechanism. Implicit artwork mentions constitute the largest category, significantly outnumbering explicit mentions, as shown in Table 7.11. The two types also differ in how often they link: explicit, named mentions resolve to Wikidata far more frequently than implicit ones (49.0 % against 30.2 % in the populated graph's clustered recognition layer, and 84.9 % against 31.9 % in the GT), so implicit mentions dominate the OOKB set (Section 8.3.4). When unique GT artwork entities are resolved against a canonical index, approximately two-thirds are out of the knowledge base rather than linked to Wikidata; and when the resolved scope widens from the stratified gold subset to the full recognition layer of the populated graph (population Run 2), the OOKB share rises further still, to 79.3 % of 15,804 artwork entities (Section 8.3.4), so the signal sharpens rather than dilutes at corpus scale. The majority of historical artwork mentions are descriptive paraphrases, and most implicit references lack any external canonical identifier. The Viewsari ontology elevates this structural absence to a first-class representation (in form of an addressable mention), ensuring that these long-tail entities persist as queryable nodes rather than being eliminated as missing data. In comparison, artists mentioned in *The Lives* reconcile to Wikidata at roughly three and a half times the rate of artworks (72.2 % against 20.7 % in the populated graph), so the canonically named population is out of the knowledge base at about a third the rate of the descriptively referenced one.

This pattern does not dissolve with a growing corpus; across all ten volumes, the populated graph carries implicit artwork mentions at nearly three times the count of explicit ones (M03, Section 9.5), so that the implicit category is not only the largest referent-introducing type in the GT but the single largest mention class in the entire graph. Following coreference resolution and entity linking in the second population run (Section 8.3.5), these mentions are consolidated into

a few thousand deduplicated artwork entities. Most of these entities do not resolve to a canonical identifier and instead persist under stable Viewsari-internal IRIs. As a result, the majority of implicitly recognized entities by ObliquER correspond to descriptive references to works absent from any existing KB.

Standard IE benchmarks assume that entities and relations are directly observable in text. However, as consolidated in several meetings with the domain experts who helped build the Viewsari framework (cf. Section 7.3.1 and Section 7.3.2), statements in Vasari's *The Lives* are interpretive, and a knowledge representation that flattens these into simple subject-predicate-object triples without epistemic considerations does not accurately represent the source.

When ObliquER identifies "a panel for the friars of Santa Croce" as an implicit artwork mention and links it to a Wikidata entity, it performs an interpretive act. In classical supervised learning, the conditions of that act were at least locatable in annotation guidelines, training datasets, and annotator records (Russell and Norvig 2010a, Pustejovsky and Stubbs 2012, Jurafsky and Martin 2026). The requirement to document interpretive decisions is not new and applies equally to human annotators. What generative models introduce is a new layer of urgency: a prompted language model constructs its interpretation dynamically, sensitive to temperature, prompt phrasing, and context window contents, with no record that divergence occurred (Meng et al. 2022, Luo and Specia 2024, Sclar et al. 2024). The widespread adoption of a small number of such models further clusters interpretive outputs around the priors of a handful of dominant decoders, so that local variability coexists with field-wide convergence (Bender et al. 2021, Mizrahi et al. 2024). Without recording the model version, prompt template, and run parameters, both layers disappear from the output. The Viewsari provenance model makes them visible: the `viewsari:promptTemplateHash` on each activity node links to the versioned template in the repository, and the `prov:SoftwareAgent` records the model identifier, so that the conditions of production stay traceable even when outputs are non-deterministic and have begun to converge at the field level.

These forms of epistemic qualification rest on a single structural decision: the model decouples the mention from its referent, anchoring each interpretive act at the level of the textual mention without presupposing an external entity to which it is resolved. The mention typology, the interpretive act typed as a `prov:Activity`, the prompt provenance recorded as `viewsari:promptTemplateHash`, and the collecting tank for OOKB entities with Viewsari-internal IRIs are the formal data-level bindings that follow from this decision. The 100 % provenance coverage result (Section 9.5.4) and the OOKB artwork entities retained in the populated KG (Table 9.21) are the empirical instantiation of this commitment: retaining OOKB entities is possible precisely because a mention can be anchored without resolving to an external

referent. The comparative modeling study (Section 9.2) confirms that none of these bindings can be represented in the variants we combined with CIDOC CRM (Doerr 2003) and its extensions.

The conceptual framework of extraction as documented interpretation holds together the ontology, pipeline, evaluation, and transferability study. The provenance gap does not merely entail a loss of detail; it results in the absence of accountability. When claims are disconnected from the circumstances of their production, both a Renaissance biographer's overheard remark and a language model's confabulation are incorporated into a KG as undifferentiated assertions, thereby transforming claims into facts. *Extraction without documentation is epistemology without accountability.* Each subsequent section addresses a specific research question within this conceptual frame.

11.2 RQ1: Representational commitments for interpretive accountability

RQ1 (Section 1.3) asks which representational commitments a KG must adopt to remain accountable to the interpretive act that produced it. We address it by designing, instantiating, and evaluating the Viewsari ontology, a provenance-aware representation that accommodates the heterogeneity of historical sources at three levels.

Answer. A KG stays accountable to the interpretation that produced it precisely when it refuses to collapse the textual mention into its presumed referent and records, on every assertion, the interpretive act that made it. Three commitments enforce this single decision: keeping mention, referent, and bibliographic source as distinct first-class entities (C1); recording the extraction activity, prompt, model, and agent on every assertion (C2/C4); and anchoring every mention to a character-offset span (the Web Annotation Ontology (Sanderson et al. 2017) layer of E1). Operationalized in the Viewsari ontology, they yield a populated graph with **100 % provenance coverage** (112,113 / 112,113 annotations), **zero unanswerable CQs** on the 70-question catalog (39 of the 52 SPARQL-translatable CQs fully answerable; the 13 partials return a structurally correct but empty result set only because their queried entity classes are not yet populated, a data-population gap rather than an ontology-expressivity gap), and a HermiT-verified consistent OWL 2 DL TBox. That the same three commitments cannot be met elsewhere is the burden of the comparative modeling study (Section 9.2), where we encoded the same Vasari passages in CIDOC-CRM alone, CIDOC-CRM + PROV-O (Lebo et al. 2013), and CRMinf.

Our observation shows that none of the three families represent all of R1–R3 natively, especially given that CIDOC CRM embodies PIEA by taking the resolved referent as its presupposed

subject; PROV-O, when attached to the referent, conflates the artwork with the claim about it, and CRMinf's own scope note disclaims the automated-extraction case. The Viewsari model represents all three, and because the CIDOC CRM family is the best-case existing candidate (Section 5.2.2.1), this elimination generalizes *a fortiori* to less expressive or more distant vocabularies; a combination of these vocabularies can be driven to the same representation, but only by reconstructing the mention/referent inversion.

The three levels provide a structural response to the three forms of heterogeneity identified in Chapter 1, as established in collaboration with domain experts. The bibliographic layer, constructed on FaBiO (Peroni and Shotton 2018), formally models the multiplicity of editions of *The Lives* and fulfills the requirements of scholarly editing as outlined in Section 2.6.1. This design ensures that every extracted entity is traceable to a specific edition, translation, volume, and paragraph, a level of granularity not achieved by any existing Vasari KG Section 5.2.4. The integration of PROV-O, in conjunction with Viewsari-internal IRIs for OOKB entities, addresses the epistemic heterogeneity of the sources, documents the conditions under which each triple is produced, and represents artworks absent from Wikidata as first-class nodes rather than as missing data.

The ontology was evaluated along three dimensions (Section 9.1.2): logical consistency, empirical adequacy via CQ coverage, and epistemic adequacy via the comparative modeling study. HermiT verifies that the synthetic ABox satisfies the axioms. The three-layer architecture and the four-way typology remain consistent under description-logic reasoning with no class rendered empty, and the most restrictive content-layer axiom, the qualified-cardinality constraint on the co-occurrence class, is met by the minimal admissible ABox. C1's separation of mention from referent is therefore not only philosophically justified but formally expressible in description logic, and the result extends to C2 and C4. We detach the reasoning chain to FaBiO and PROV-O prior to classification because both ontologies include SWRL rules that HermiT cannot process. As a result, the fully populated 2.45-million-triple graph is not classified by any reasoner. This limitation does not diminish the consistency claim. The synthetic ABox is explicitly constructed to instantiate every class and exercise every object property at least once, thereby spanning the full axiomatic surface of the schema. The populated graph introduces no additional classes, properties, or axioms; rather, it instantiates these patterns at scale. Consequently, the consistency claim is bounded rather than weakened. HermiT verifies schema-level (TBox) consistency and class satisfiability, and this guarantee extends to the populated graph, since it adds only assertions over an already-verified vocabulary. However, the reasoner check does not establish assertion-level constraint satisfaction at scale: a population run could violate a data-level axiom, such as the qualified-cardinality constraint on the co-occurrence class, which would manifest as a data-quality defect of that run rather than as a schema inconsistency. We therefore assess full-graph behavior through CQ coverage (see Section 9.1.3) rather than description-logic entailment. This

approach aligns with C3's distinction between hermeneutic and mechanistic provenance. In the context of a humanities knowledge graph, the central accountability question is not whether a reasoner can infer every implicit subclass relation, but whether the construction process remains queryable in terms relevant to scholarly inquiry. Moreover, using entailment to derive new facts violates our theses in C1 and C3. An inferred triple is itself an interpretive act that would require its own provenance. The ontology's predominant purpose is to fix a shared vocabulary and target schema that constrain generative extraction to a typed, queryable structure, making the C1 mention/referent separation and the C2 provenance edge structural rather than advisory and rendering every claim contestable through SPARQL and linkable to Wikidata and FactGrid (Simons 2024); reasoning as a documented activity in its own right (Section 12.3) is an additional return.

Beyond consistency, the ontology's empirical adequacy is borne out by the CQ coverage reported in Section 9.1.3 and summarized in Table 9.1: 39 fully answerable, 13 partial, 0 unanswerable, and 18 descriptive out of a 70-question catalog turn the three representational commitments into queries a scholarly reader can run against the endpoint. The closure rules of Section 9.1.3 reconstruct the per-instance anchor edges that five Phase-I CQs require without introducing new classes, properties, or axioms, materializing them through the `prov:Activity` chain reified by COMP-PROV. All 13 partial cases correspond to declared but unpopulated entity classes, confirming a data-population reading over a structural one, while the 18 descriptive CQs provide the empirical foundation for C3, with their answers located primarily in the deployed endpoint.

Answerability does not, by itself, show that the answers are grounded in the populated data, which the provenance-transparency evaluation in Section 9.5.4 establishes using two metrics, both computed on the canonical Run 2 closed graph (Section 8.3). The principal one is 100 % coverage: every annotation is linked via a `prov:wasGeneratedBy` chain to both a typed activity and a software agent, whose inventory is reported in Section 8.3.3. This operationalizes C2 and contrasts with the 0 % of the baseline Vasari KG (Santini et al. 2022, 2024). The figure is not a schema tautology but emerges from the ontology and the extraction pipeline jointly, establishing provenance as a constitutive property of scholarly assertion under generative extraction at scale rather than an auxiliary addition.

Coverage establishes that every annotation has a documented origin but not which kind of agent supplied it, and it is this distinction that the strict ratio M07 draws by partitioning the provenance chains by agent type, evaluating to 0.751 (75.1 %). The 24.9 % shortfall is not a population failure but a consequence of the criterion's `prov:SoftwareAgent` requirement. Of the 27,898 annotations outside the strict set, 2,439 are the manually annotated GT mentions, which carry a full activity chain attributed to a human annotator (`prov:Person`) rather than a `prov:SoftwareAgent`, and the remaining 25,459 are non-mention annotation nodes that carry

neither the direct in-paragraph edge nor a software-agent activity. Every typed mention, including all 22,283 coreferents, carries the direct edge, so the strict ratio isolates the machine-extracted, paragraph-anchored fraction and excludes the human-annotated gold layer by agent type rather than for want of provenance. This is what C3 and C4 require, since the model records not merely *that* every annotation has an origin (M06 already establishes universal coverage) but *which kind of agent* produced it, so human and machine interpretation stay distinguishable, and the partition is recoverable by a single filter on agent type, possible only because COMP-PROV reified the agent as a typed node. The same separation explains the baseline contrast. The 0 % Vasari KG returns nothing to a scholar asking how a co-occurrence was produced, whereas the Viewsari graph returns the paragraph, run, model, and prompt hash, which is what it means for hermeneutic provenance to be implementable as data.

While consistency, answerability, and provenance characterize what the commitments do in principle, the three population metrics of Section 9.5 establish that they hold across a graph of this size. The realized 9.7 triples per mention (Section 9.5.1) sit well below the 12–18-triple budget and are dominated by the Web Annotation Ontology and PROV-O slots any annotation-plus-provenance representation requires, so C1's first-class mentions and C4's COMP-PROV pattern impose no verbosity cost the graph fails to absorb. The 72.2 % person-linkage against 20.7 % for artworks (Section 9.5.2, Table 9.21) is the asymmetry the long-tail argument predicts (Section 7.4.9.4). Wikidata's Renaissance coverage is densest around artists, while the works trail into a long tail, and the 12,529 OOKB artworks retained under stable internal IRIs (Section 8.3.4) survive as queryable, fully provenanced nodes that a surface form linker would silently drop. The 100 % volume-and-biography and 92.3 % page-level reachability (Section 9.5.3) materialize the FRBR chain to the manifestation level, so a reader can follow a provenance chain from a co-occurrence back to the de Vere paragraph that sourced it. These metrics close the loop. Reasoner-level coherence, catalog-level answerability, schema-and-population provenance, and population-scale operational properties all converge on the same RQ1 answer.

Verdict against the answerability criterion. The criterion set for RQ1 in Section 1.3 is met. The three required commitments are named (C1, C2/C4, and span anchoring), shown unmet by the CIDOC CRM family in the comparative modeling study (Section 9.2), and operationalized in an ontology whose schema HermiT verifies as consistent with all classes satisfiable under OWL 2 DL (Section 9.1.2), that is answerable against the seventy CQs with none unanswerable (Section 9.1.3), and that is provenance-complete on the populated graph (Section 9.5.4). Full-graph classification is infeasible at the 2.45M-triple scale and is not part of the criterion; the populated graph introduces no terminological construct beyond those the verified schema already covers. The representational-commitments question is thereby answered on the terms set for it in advance.

11.3 RQ2: Large language models as interpretive agents for implicit and out-of-knowledge-base entities

RQ2 (Section 1.3) asks under what conditions LLMs can recognize and link the implicit and OOKB mentions that surface form-level baselines miss. We answer RQ2 through the ObliquER pipeline, demonstrating that LLM-based generative extraction can identify entities that are invisible to conventional NER systems while maintaining the provenance commitments established by RQ1.

Answer. An LLM recovers the implicit, and OOKB mentions that surface form-level baselines miss when it is constrained to act as an ontology-typed interpretive agent. On the criterion that measures whether the right entities are found, ObliquER recognizes Viewsari artwork mentions at **Partial F1 = 0.697** and **Type F1 = 0.642** under a modular, ontology-typed prompt and a retrieve-then-rank linking stack; the lower exact-span figure (0.334) reflects boundary subjectivity that even two human annotators exhibit (exact-span agreement 0.43 against 0.83 partial, Section 9.4.1), not a failure to locate the entity. Against the surface form-only baselines the difference is categorical rather than gradual, and is read off the strict criterion: ObliquER reaches implicit **Strict F1 = 0.321** where UniNER-7B (Zhou et al. 2024c) and GLiNER2 (Zaratiana et al. 2025) score **0.00–0.02** because they are, even if in principle capable, unable to locate implicit spans at all, and it links implicit mentions at **KB-linked F1 = 0.385** against **0.039 / 0.059 / 0.028** for ReFinED (Ayoola et al. 2022) / BLINK (Wu et al. 2020) / GENRE (De Cao et al. 2021). These enabling conditions are architectural rather than parametric, as we isolate in the ablation. Stripping the domain-specific content layer while holding the modular pipeline identical (ablation A2, Table 9.17) collapses Strict F1 by **25 %** on `gpt-oss-120b` and **38 %** on `gemma-4-31b-it` once the model is returned to a generic NER prompt. Within that content layer, the ablation attributes **72–95 %** of the contribution to the ontology-typed definitions rather than the demonstrations, the parametric-only linking floor collapses to **0.000** KB-linked F1, and the recognition surface is invariant across decoders (cross-model Exact-F1 envelope **0.048**; zero-shot gap **0.007**).

ObliquER's principal contribution lies in its ability to recognize implicit mentions of artworks, the descriptive expressions that refer to specific, identifiable artworks without recourse to canonical names. In Chapter 1, we have outlined this conceptual line, where Cavazza and Zweigenbaum (1992) already argued that recovering implicit content requires constructing a model of the described situation rather than extracting lexical triggers alone; Perera (2016) distinguished implicit entities (no referring expression, real-world referent that must be inferred) from coreferents (referring expression present, no new referent introduced); and Hosseini (2019) formalized the

problem in the contemporary IE context. The case for ObliquER is that the standard surface form-level classifiers and LLMs (UniNER-7B, GLiNER) achieve 0 % recall on this mention class (they can only classify and identify named entities of a predefined type), so any non-zero implicit recall is not an incremental gain over the baseline but a qualitative extension of the recognition surface. ObliquER achieves an implicit Strict F1 of 0.321 on Viewsari, compared to baseline values of 0.00 to 0.02 (see Table 9.6). This result empirically substantiates the theoretical trajectory established by Cavazza and Zweigenbaum (1992), Perera (2016), and Hosseini (2019). Implicit recognition is not only conceptually coherent but also operationalizable. Crucially, this is achieved through a typology of mentions. The methodological advance consists of reframing implicit reference as a problem of typological identification. This shift expands the recognition surface and aligns with the interpretive framework articulated in the introduction.

Although the Strict F1 score of 0.321 may initially signal a weak benchmark result, we emphasize that this is not the case, given the 0.00 to 0.02 range established by state-of-the-art baselines. This low baseline is determined by architectural constraints rather than by parameter tuning and specifically affects the descriptive references prevalent in Vasari's prose and identified by the canonization signature as culturally significant. In this context, ObliquER's 0.321 is the first nonzero measurement for a recognition surface that surface-form-only baselines cannot reach (even if they could, theoretically). Two effects compound here and are worth separating. Using a generative reader at all already lifts implicit recognition off the surface form-level architectural zero, while the ontology-typed framing supplies the larger, isolable share of the gain. The generic-prompt ablation A2 still scores 0.249 Exact F1 on the same decoder (Section 9.4.6), so the rise to 0.321 is not created by ontology-typing alone but is attributed to it only in the part the ablation isolates. It should be interpreted as the error rate of a newly developed capability applied to a particularly challenging text, rather than as the accuracy of a fact-retrieval system aiming for a high benchmark score. When evaluated using boundary-tolerant criteria, this capability achieves a Partial F1 of 0.697 and a Type F1 of 0.642. The exact-span metric is reduced by the same boundary subjectivity that limits human inter-annotator agreement to 0.43 exact versus 0.83 partial (Section 9.4.1). Therefore, the strict score understates the system's capability and is most useful as a comparative baseline. The primary contribution of this dissertation is not the 0.321 score itself, but rather the modeling approach that enables and substantiates this result. This includes the mention typology, which allows a generative model to produce implicit predictions (C1, E1), and the COMP-PROV provenance layer, which documents the conditions under which these predictions are made (C2–C4, E2–E3). The reported score serves as evidence of the modeling's effectiveness; it is not the central focus of the work.

This advantage persists across the entire set of mentions. The descriptive comparison of Section 9.4.3 reports an ObliquER span-based Strict F1 of 0.334 on Viewsari (All), compared to 0.09 for the strongest span-only baseline, GLiNER2. On the Explicit subset, ObliquER outperforms

GLiNER2 by approximately 14 points (0.375 versus 0.23). The remaining 10-point difference in the All subset arises from the implicit mentions, which surface form-only baselines are structurally unable to predict. Implicit-typed gold mentions are typically descriptive paraphrases (*"a panel for the friars of Santa Croce"*) rather than canonical titles (*"the Last Supper"*). The mention typology thus provides the categorical mechanism by which a generative model can produce implicit predictions, shifting the evaluation from explicit recognition to a distinct recognition surface. The same error pattern appears in ArtPedia, where implicit-to-Explicit confusions predominate, as in Viewsari, but with a substantially larger implicit subset (see Section 9.4.7). This pattern demonstrates that the boundary between descriptive paraphrase and canonical title is determined at the corpus level, not by model behavior. The modest decrease in performance from Viewsari to ArtPedia ($0.334 \rightarrow 0.319$) reflects this corpus shift rather than a failure of the extraction pipeline. ObliquER continues to outperform surface form-only baselines under equivalent conditions; the quantitative analysis of this claim is addressed in Section 9.4.7. The three principal NER failure modes identified in Section 9.4.3 – confusion between descriptive mentions and generic artistic descriptions, errors in multi-clause implicit-span boundary detection, and type misclassification – partition the residual error. Each maps to a specific remediation strategy, namely refining the negative clause in the entity-definition block, extending few-shot demonstrations to longer descriptive spans, and enhancing annotation guidelines with explicit boundary heuristics.

We compared our canon prompt template used for the evaluation (ontology-guided) against other schemas; the CoT variant extends its architecture with auditable reasoning traces stored as provenance metadata. The prompt-variant comparison of Section 9.4.6 aims to answer where the CoT variant consistently improves extraction accuracy over the ontology-guided variant. Under `openai/gpt-oss-120b`, it returns span-based Strict F1 $= 0.322$ against the production variant's 0.334, a 0.012 deficit that sits inside the bootstrap of either configuration, so CoT is not an accuracy-improving prompt structure on Viewsari, as has been established in reference literature like Winkle et al. (2026).

We also examined whether the prompt architecture carries the recognition gain by comparing changes to the decoder. The cross-model evaluation in Section 9.4.5 runs the production ontology-guided recognition prompt through `google/gemma-4-31b-it`, with all other components held constant. ObliquER's recognition behavior with the template is stable across model changes, rather than a property of any specific inference target. The A3-equivalent zero-shot row returns span-based Strict F1 $= 0.330$ on `gpt-oss-120b` and 0.323 on `gemma-4-31b-it`.

This finding substantiates our empirical claim that, within ObliquER, the ontology-typed prompt architecture enables generative models to recover implicit references that surface form-only baselines cannot capture, thereby facilitating transferability across decoders. The residual Type-minus-Strict gap is attributable primarily to intrinsic span ambiguity rather than to limitations in

decoder capacity, which aligns with the boundary-detection perspective outlined in Section 9.4.3. The few-shot count, correspondingly, does not close this gap, which is why the production 2-shot setting was retained. The same invariance also has a theoretical reading. Under the canon-convergence thesis of Section 11.1, cross-model convergence on a recognition task is exactly what would be predicted. Cross-model consistency is precisely the signature a prompt-driven architecture predicts: if the ontology-typed prompt carries the representational work, decoder identity should be a secondary variable – and the convergence of two distinct models confirms this. Under our canon-convergence thesis ('winners writing history'), this consistency has a second reading, namely that foundation models trained on overlapping corpora will reproduce the same canonical biases, making cross-model agreement on implicit references a provenance signal rather than a validation of recall. The Viewsari response is to record the conditions of production in prompts and agents, so that when convergence becomes a methodological provenance trail, it becomes visible at the data level.

We further isolated our components with an architectural ablation in Section 9.4.6. Table 9.17 decomposes the recognition F1 contribution into three pillars under both decoders. The no-domain condition (A2) is the controlled counterfactual for the architectural-not-parametric claim of the RQ2 answer, as it removes the expert role, the ontology-typed entity definition, and the special-cases block while leaving the pipeline staging untouched, returning the model to a generic NER prompt that must rely on its parametric priors. A2 is accordingly the harshest single ablation in the table, dropping span-based Strict F1 by **25 %** on `gpt-oss-120b` ($0.334 \rightarrow 0.249$) and **38 %** on `gemma-4-31b-it` ($0.361 \rightarrow 0.224$), so the recovery documented above is an effect of the ontology-typed framing and not of what either decoder already knows. Under GPT the ontology-typed content layer alone (A3 vs A2) accounts for $+0.081$ span-based Strict F1, the pipeline staging (production vs A1 monolith) for $+0.050$, and the two annotated demonstrations (production vs A3) for $+0.004$, a margin well within the per-paragraph bootstrap interval of Section 9.4.1; under Gemma the same decomposition gives ontology $+0.099$, staging $+0.061$, demonstrations $+0.038$. Jointly with the cross-model evidence, these variants bracket the recognition contribution from opposite directions. A2 holds the decoder, and with it the parametric store, fixed and varies only the prompt, so its 25–38 % collapse cannot be charged to anything the decoder knows; the cross-model run instead holds the prompt fixed and swaps the decoder, and the ontology-only zero-shot rows still agree to 0.007 span-based Strict F1 (Section 9.4.5). The within-decoder arm is the decisive one, since it directly controls the parametric store, while the cross-decoder arm corroborates it, subject to the caveat of shared provenance noted above. Together, they locate the capability in the ontology-typed prompt structure, not the demonstration budget and not the choice of decoder, with the cross-decoder discrepancy in the demonstration sub-component, the one place where decoder identity shifts the ranking quantitatively but not qualitatively.

These chapter-wide figures average over substantial variance across the 16 Viewsari biographies (Section 9.4.3, Table 9.10). Fifteen sit in a tight band; Michelangelo Buonarroti is the lone outlier, at Type F1 0.437, well below both the next-lowest biography and the chapter-wide aggregate. Where the deficit falls matters more than its size. The Methodology (Section 7.4.4.2) names Michelangelo's biography as the anchor case for prompt and schema development, retaining more annotated paragraphs than the cap applied to every other biography, so the biography with the most design-time influence on the production prompt is also the one it performs worst on at evaluation. Read against the central claim about making the construction process visible, the Michelangelo outlier is the signature of within-corpus interpretive heterogeneity. Vasari devotes more text to Michelangelo than to any other artist, and that prose is the densest in descriptive iconographic detail, so the epistemic uncertainty that Chapter 2 names as constitutive of humanities sources is distributed unevenly across the biographies rather than flat across the corpus. The most canonized artist in *The Lives*, carrying its largest gold-mention count, is the one ObliquER finds hardest, plausibly because canonization multiplies the implicit, descriptive, iconographically overlapping references that resist surface recognition; on this reading, the outlier confirms the thesis rather than contradicting it. The reading does not displace the design-time confound; it cannot separate from Michelangelo's biography, which is at once the development anchor on which the schema was tuned and the densest canonical-reference network in the corpus, so the claim is advanced as a plausible reading, not a demonstrated cause. Either way, the distribution becomes queryable. Per-biography Type F1 is recoverable from the populated graph through the bibliographic-layer chain and the mention-level provenance edges that tie each annotation to its paragraph, biography, volume, and translation, so the heterogeneity is something the ontology was built to surface rather than only a row in a table.

At the linking stage, the pipeline encounters the characteristic long tail of CH data. About two-thirds of the manually annotated gold artwork entities are absent from Wikidata (63.4 %), and across the full populated graph the share is higher still, at 79.3 % of 15,804 deduplicated artwork entities (population Run 2), so widening the resolved scope from the stratified subset to the full recognition layer makes the long tail more pronounced rather than less (see Section 8.2.2). This distribution reflects an inherent property of the data rather than a performance limitation. Vasari documents numerous works that lack Wikidata entries because they no longer exist, were never individually cataloged, or do not meet the platform's implicit inclusion criteria. The pipeline addresses this by elevating OOKB detection to a primary output. The retrieve-then-rank candidate generation process provides an LLM ranker that returns either a Wikidata identifier or an explicit OOKB designation. OOKB entities are maintained under stable, provenance-aware Viewsari-internal IRIs, ensuring their availability for future reconciliation as coverage expands. This operationalizes the long-tail mechanism discussed in Section 11.1 at the linking stage. The

works most central to Vasari's canon frequently lack canonical identifiers, and any architecture that omits them would fail to capture the dynamics of canonization central to this thesis.

The standalone EL comparison (see Section 9.4.4) establishes that ObliquER outperforms existing baselines, with the most pronounced improvement on the implicit subset. Candidate-recall metrics account for this result because surface form linkers seldom retrieve the correct implicit artwork given that descriptive paraphrases are absent from indexed dictionaries. In long-tail entity scenarios, both ObliquER and the strongest baseline (ReFinED (Ayoola et al. 2022)) achieve comparable performance on the OOKB category, but through fundamentally different mechanisms. The baseline fails when the candidate set is empty, whereas ObliquER assigns the OOKB label through an explicit ranking decision. This distinction is significant. In ObliquER, OOKB assignment is a deliberate outcome of the model's inference process. On the explicit subset, where an incorrect OOKB assignment constitutes a true linking error rather than a limitation of long-tail coverage, the systems' performance diverges. The ArtPedia comparison (Table 9.12) corroborates this pattern with an inverted error profile: explicit linking for Wikipedia-grounded artworks saturates across all systems, while on the implicit subset, ObliquER achieves a KB-linked F1 of 0.713 (compared to 0.385 for Viewsari), surpassing the next-best baseline by a substantial margin. The advantage is thus concentrated on the implicit-linking task, while explicit-linking performance reaches a ceiling, since baseline linkers already resolve these cases. Two control experiments isolate the source of this improvement. When the model must produce a Wikidata identifier from parametric memory alone, the monolithic variant yields zero linked F1, despite generating spans (Table 9.15), indicating that the retrieve-then-rank architecture, not pre-training, is responsible for the observed gains. Additionally, rewriting the raw mention into the LLM-generated structured metadata record (Section 9.4.4) improves candidate recall across all retrievers and subsets, confirming the necessity of the query-rewriting stage. Each per-mention decision is recorded in the KG as a queryable provenance entry.

ObliquER contributes 84,215 typed mention annotations to the canonical Run 2 KG; together with the 2,439 manually annotated ground-truth mentions this yields 86,654 typed mention annotations in total, over 15,804 unique artwork entities (3,275 Wikidata-linked, 12,529 OOKB), 443 person entities (320 Wikidata-matched), and 632 co-occurrences.

Verdict against the answerability criterion. The criterion for RQ2 in Section 1.3 is satisfied. The documented pipeline identifies implicit, and OOKB mentions with Partial F1 of 0.697 and Type F1 of 0.642, metrics that directly reflect entity identification accuracy. On the strict criterion used for baseline comparison, the pipeline achieves rates that surface form-level baselines cannot attain by design (implicit Strict F1 0.321 versus 0.00–0.02; implicit KB-linked F1

0.385 versus a maximum of 0.06). Ablation analysis attributes this improvement to the ontology-typed prompt, not to incidental tuning (accounting for 72–95% of the content-layer contribution; parametric-only linking yields a floor of 0.000). Recognition performance remains stable across decoders. The evaluation thus establishes, rather than merely asserts, the conditions under which LLMs, when treated as interpretive agents, recover this recognition surface.

11.4 RQ3: Design patterns and methodological transferability

RQ3 (Section 1.3) asks what reusable ODP emerges from modeling extraction as a documented, provenance-aware interpretive process, and how that foundation transfers to structurally distinct corpora. In effect, we ask whether the Viewsari methodology is specific to Vasari's *The Lives* or addresses challenges common to historical KG construction more broadly; the evidence from Chapter 10 supports the latter.

Answer. The central reusable pattern is COMP-PROV (Section 10.3.2), the attribution of an interpretation to a parametrized machine process, and it transfers to a structurally distinct corpus by holding the ontology core fixed (extending it to fit the pilot) and changes in the prompt content for the ObliquER pipeline, with the transfer bounded to recognition and representation rather than end-to-end linking. The supporting evidence is reported in full in Chapter 10: across ICON (Sartini et al. 2023), Polifonia (de Berardinis et al. 2023), ODEUROPA (Lisena et al. 2022), and Viewsari, only PROV is shared (Table 10.3), so COMP-PROV is a genuine addition to the existing DH pattern repertoire rather than a renaming of one. The Kloß *Annalen* pilot reapplies the pipeline across five axes (language, period, entity type, target knowledge base, referential strategy) through the content layer alone, with recognition carrying over and the coded alias, the German conspiratorial analog of the Vasari implicit mention, the highest-scoring populated type (Section 10.2.2). What this section adds to those results is their reading in light of the dissertation's central claim. We generated a mock-KG to show that the Viewsari ontology can transparently carry the process of LLM-mediated interpretation into a graph for a domain it was never designed for. As a Yin-style theoretical replication rather than a statistical generalization across a population of sources (Section 10.1), the pilot is primarily a test of the harmony between the two transferable components of ontology and ObliquER. This transfers further into the abstract COMP-PROV pattern.

The pilot also operationalizes RQ3 back to the central claim. Because FactGrid cites the *Annalen* as a source for many of the entities the text mentions, the same source passage can carry an

annotation from a human scholar (the FactGrid editor citing Kloß § 174 as evidence for Knigge's lodge visit) and from a parametrized machine agent (an ObliquER run emitting a typed mention with extraction provenance), and the two coexist in the graph without competing for the same slot (Figure 10.6). The human-curated claim and the machine-extracted mention agree on the referent and differ only in the provenance they carry, so each extends the other. The pre-existing FactGrid claims serve as a partial silver standard against which the machine annotations can be fact-checked, while the machine layer supplies the evidence span, prompt configuration, and run identifier that the FactGrid reference mechanism requests but cannot itself record. A graph that hosts both is the operational form of the dissertation's claim that interpretation in the age of generative AI requires that the construction process be visible from *all* sides. The coded alias transfers the implicit reference-capability, not an equated difficulty. It is a single opaque token ("Philo", "Minos") rather than the multi-clause descriptive paraphrase of a Vasari implicit mention ("a panel for the friars of Santa Croce"), and part of its score reflects aliases the prompt itself decodes, so the coded-alias figure should be read as the Kloß analog being easier, not as the descriptive-reconstruction capability transferring at higher accuracy. The pilot stops before the step that turns recognition into a graph. The FactGrid entity-linking stage, ObliquER's defining contribution exercised in a new domain, is demonstrated only as the serialization of a human and silver GT into the ontology, not evaluated end-to-end.

Furthermore, the comparative analysis (Section 10.3) and the transfer pilot together yield a reusable bundle of abstractions for DH KGs. The Annotation pattern (ANN), parent of the Co-occurrence pattern (Section 10.3.3) and the COMP-PROV (Section 10.3.2). None of the four comparison projects instantiate all three; Viewsari is the first to combine them, and that joint instantiation enables the federated, double-view scenarios the pilot demonstrates. The same property makes the contribution adoptable beyond Viewsari. COMP-PROV is intentionally lightweight, so that a typed software agent and a reified extraction run layered onto a standard PROV-O chain. The distinctive feature is that provenance attaches to the annotation rather than to the referent. It therefore applies directly to any DH graph that already pairs PROV-O with an annotation or mention layer over extracted text, as both ODEUROPA (Lisena et al. 2022) and the MEETUPs KG as part of Polifonia (Polifonia Project 2023) do, and the underlying commitment carries into neighboring text-based disciplines (Section 11.8).

Verdict against the answerability criterion. The criterion set for RQ3 in Section 1.3 is met. The comparative analysis isolates COMP-PROV, a reusable pattern unique to generative extraction and absent from the three comparison projects, and the Kloß pilot shows it transferring to a structurally distinct corpus through the prompt content layer alone, recognition carrying over at span-based Strict F1 0.600 / 0.683 (coded-alias 0.828 / 0.878).

We articulate the transferability claim along two methodologically distinct axes, each requiring separate consideration. The first axis, represented by the ArtPedia comparison, holds the extraction task constant while varying only the corpus. Specifically, we apply the production ObliquER pipeline to a period- and register-shifted English-language corpus that exhibits a different implicit-reference profile but retains the same Wikidata target (Section 9.4.7). This experiment yields a KB-linked F1 of 0.733 overall and 0.713 on the implicit subset, compared to 0.004 for the strongest baseline. These results quantitatively confirm that the ontology-typed pipeline and its linking stack transfer effectively across corpora that share the same target knowledge base.

In contrast, the Kloß/FactGrid pilot described in Section 10.2 introduces simultaneous shifts in language, century, entity type, target knowledge base, and referential strategy. The objective here is not to replicate the ArtPedia benchmark but rather to demonstrate that the Viewsari ontology and the COMP-PROV pattern transfer solely through the prompt content layer. Accordingly, we do not evaluate end-to-end entity linking against FactGrid in this pilot. Instead, the contribution lies in demonstrating that human-curated provenance and machine-extracted provenance coexist within the same graph and under the same schema (Figure 10.6). This structural property of the pattern cannot be captured by any F1 metric. This constitutes a Yin-style theoretical replication (Yin 2018) rather than a statistical generalization across a population of sources, bounded to recognition and representation by design (Section 10.1).

Taken together, these two axes delimit the generalization claim from complementary directions. The ArtPedia experiment confirms quantitative transferability within a shared knowledge base while holding the extraction task fixed. In parallel, the Kloß/FactGrid pilot confirms structural transferability across five dimensions of corpus shift, with the ontology and pipeline held constant. Thus, the criterion established for RQ3 in Section 1.3 is satisfied along both axes. The scope boundary of the far-domain pilot – which demonstrates recognition and representation while reserving end-to-end out-of-domain linking for future work (Section 12.3) – reflects the experimental design rather than a deficiency in the evidence.

11.5 Complementary forms of interpretation: CRMinf and Viewsari

Chapter 9 establishes that CRMinf and Viewsari differ (four scope axes: the subject of representation, the agent, the epistemic attitude of belief, and the provenance of the generative run), also given that CRMinf's specification excludes automated extraction interpretation. Our showcase of the Bologna-Nero walkthrough cannot be transposed to an ObliquER-produced annotation without violating each of these axes (Section 9.2.4). Read against the central claim of this

dissertation, that divergence shows the categorical difference of how CRMinf models scholarly arguments about the world, namely who reasoned over which sources, with which inference logic, to arrive at which belief about which referent; Viewsari models computational annotations on text, namely which span of which paragraph of which edition was typed as which kind of mention by which extraction run. The de facto standard for modeling interpretation in DH and CH use cases, also shown in other KGs like the Vespasiano da Bisticci's letter edition by Tomasi (2013), does thus not hold when the goal is to represent a claim about textual mentions based on an extraction process (Chapter 3).

The natural sisterhood between these two lies in their focus on *interpretation*. Our central claim, for interpretation done by generative AI as an epistemic act (Section 11.1), accepts this premise. What differs is not whether interpretation occurs but its form. Viewsari is the counterpart to what CRMinf tackles, recording the externally observable conditions of the activity (prompt template, model identifier, run timestamp, sampling parameters, and reasoning trace) as queryable data, in place of an interior that cannot be reconstructed, especially given that we cannot model 'beliefs' of LLMs, as requested by CRMinf. If extraction from interpretive sources is itself an interpretive act, the rigor CRMinf already grants a scholar's argumentation has no principled reason to be withheld from its computational counterpart, and Viewsari is the disciplinary follow-through that supplies it under the vocabulary that the shape of interpretation requires. The ontologies, as complements, are rigorous in their own scope. Viewsari surfaces every mention together with the external provenance of the extraction that produced it, adding a further layer for clustering and linking these textual mentions to a potential real-world referent, which lies outside what CRMinf set out to model, while Viewsari has no class for a scholar's adopted belief or a provenance assessment of a transmitted source. A combination of both in one graph is thus not out of scope, and adding a CIDOC layer where Viewsari models real-world referents for extracted artworks could further be supplied with interpretations done by human scholars and their held beliefs could be modeled in CRMinf.

This combination also locates Viewsari in a definite place within the larger documentation stack, and that placement is itself a conceptual contribution of the dissertation. CIDOC CRM and the event-centric vocabularies layered on top of it presuppose that the entity under discussion is already a known referent, so that a production event, an attribution, or an exhibition history can be hung on a stable node. Viewsari works one step earlier, directly on the text. From a character span, it first decides whether the span denotes an artwork at all, which is the entity-recognition question, and then which artwork it denotes, which is the entity-linking question, and only the resolved referent of that second step is the node a CIDOC layer would subsequently elaborate. The two vocabularies meet at the entity-linking output, the Wikidata QID or the minted OOKB IRI, so that Viewsari supplies the upstream evidence (the span, the mention type, the extraction run) for a referent that downstream event modeling then treats as given.

Seen this way, Viewsari occupies on the machine-extraction side the position that ontologies such as HiCO (Daquino and Tomasi 2015) occupy on the human-interpretation side. HiCO records the interpretive act by which a scholar arrives at a claim about a work; Viewsari records the interpretive act by which a parametrized run arrives at a typed mention and its candidate referent. A useful image is of a reader working through *The Lives* with a highlighter, who first marks a stretch of text as naming a painting, then researches which painting it is, takes notes, and cites a source. The highlighting and the identification are already interpretive moves on text, prior to any statement about the painting as an object in the world, and Viewsari is the vocabulary that makes those two moves explicit and queryable rather than leaving them implicit in the eventual referent.

This closes the loop to the central claim and the theses of Chapter 1. We set the boundary of representation before the pre-interpretive entity assumption, because the judgment that a span "is an artwork" is not a neutral preliminary to interpretation but an interpretive act in its own right, as is the judgment of which artwork it is. CIDOC and CRMinf begin once a referent is fixed; Viewsari is what models the recognition and linking that fix it, and it does so for the machine extractor as deliberately as CRMinf does for the human scholar. The framework therefore does not compete with the established cultural-heritage stack but supplies the layer beneath it that the generative-extraction setting makes newly necessary.

11.6 On agentic large language models and the future of Viewsari

The growing spurt of agentic systems in LLMs, which ground their outputs through tool use, retrieval, and CoT traces, adds another layer of implicit accountability to the use of generative systems for knowledge production. While seemingly transparent with a tool-augmented agentic loop, this does not change *what* must be represented in the subsequent KG, not *whether* it must be. An agentic extraction does more, not fewer, things, so the operative question is not whether agentic LLMs render Viewsari obsolete but how Viewsari extends to represent the additional artifacts they produce. Viewsari is not a confidence-improvement device that marks an extraction as reliable; its purpose is to make every commitment of the extraction process representable in the resulting graph, which is the dissertation's central claim (Section 11.1).

That extension is already underwritten by the results of this dissertation. The Viewsari design treats the `prov:Activity` of an extraction as a first-class node that carries the model identifier and prompt-template hash, with full provenance coverage reported in Section 9.3 and formalized as COMP-PROV (Section 10.3.2). Even the latest trace-as-grounding designs such as the AEVS pipeline of Yang et al. (2026), which ties each generated element to an identified source span,

supply the textual position of an anchor rather than the interpretive convention under which the model decided the anchor denotes what it denotes (Chapter 2); such staging multiplies the artifacts a provenance ontology must capture, not the reasons to dispense with one.

This also sharpens the boundary drawn in Section 11.5: an agentic system that retrieves and self-justifies is not thereby brought closer to CRMinf's human-actor ontology, since the retrieval, reasoning, and justification are still produced generatively, a process to which no honesty precondition attaches (Plaat et al. 2025, Liu and Shu 2026). Agentic extraction therefore strengthens rather than weakens the case for Viewsari's vocabulary. The ontology is the kind of artifact whose relevance grows with the complexity of the construction process it models, and a richer process is more to represent, not less.

11.7 Building a knowledge graph off of an 'unreliable' system

Before turning to methodological reflections, we address the elephant in the room – *the source is contested, the extraction is imperfect, the model is a black box, therefore the representation is unreliable*. The implicit conclusion is that provenance-aware KGs of interpretive sources are an unstable foundation for scholarly use, and that the appropriate response is either to restrict KG construction to domains where facts are stable and extraction is accurate or to defer humanities KG projects until generative extraction reaches a higher accuracy threshold (Piotrowski 2019, 2026).

We argue, however, that the inference runs in the opposite direction. It is precisely because sources are contested, extraction is imperfect, and models are opaque that provenance must be a first-class element of the representation. A KG that refuses to document these conditions does not solve the problem; it conceals it. The conventional fact-oriented KG, applied to Vasari's *The Lives*, would not eliminate interpretive uncertainty; it would simply suppress the documentation that allows readers to recognize where the uncertainty lives. Viewsari foregrounds what such a graph would suppress. Our argument rests on the distinction Chapter 2 draws between *statistical* and *epistemic* uncertainty (Mugleston et al. 2025, Aroyo and Welty 2015, Plank 2022), the difference between computational models, which record uncertainty as probabilities or confidence scores, and interpretive sources, which surface a different kind of uncertainty. Structural ambiguity in language, conflicting evidence, and changing scholarly views that no probability-over-labels can express. Historical scholarship has long acknowledged uncertainty as missing, incomplete, contested, or ambiguous information; what Piotrowski (2019) and Piotrowski (2026) argue, and what the dissertation operationalizes, is that in the DH context this implicit uncertainty (hidden in

the extractions themselves) must be made explicit and formally represented so that computational processing and reasoning can carry it forward rather than flattening it. The unreliability objection presupposes a uniform notion of reliability against which all representations are scored; the uncertainty taxonomy of Chapter 2 shows that the objection's notion is the statistical one and the dissertation's commitment is to the epistemic one.

Where such an objection treats source contestation, extraction imperfection, and model opacity as conditions that disqualify a representation, the dissertation treats them as conditions that determine the form a representation must take. The 0.321 Strict F1 of ObliquER on implicit artwork extraction and the 79.3 % OOKB rate against Wikidata in the deployed KG are not defects that the Viewsari representation hides but instead foregrounds. The representation is reliable in a different sense, which is not statistically reliable as a record of facts, but epistemically reliable as a record of how each claim came to be made. This inversion is, in the end, historical source criticism carried over to the machine (Bernheim 1908, Schlotheuber and Bösch 2015, Fickers 2020). Applied source criticism requires that a source be documented well enough to be criticized, with its origin, agency, and the conditions of its making recorded so that later readers can build upon it. Viewsari asks exactly that and requires not that the model be trustworthy, but that the conditions under which it produced each claim be documented well enough that the claim can be subjected to the same critical reading a historian would bring to any other source.

With this explanation, we address recurring objections such as "how reliable is the extraction, really?" or, as art historians often put it, "but how are we to believe this output?" The honest answer is that with ObliquER, we do not claim to verify each contested attribution Vasari makes in his work. This is precisely why the provenance trail, rather than an accuracy figure, is the operative response. It does not assert that an extraction is correct; it makes visible where it came from, naming the paragraph, the edition and translation, and the model and prompt, so that the question of reliability is relocated from the output to the source, where a reader can adjudicate it. This is also why the art historians who shaped the requirements did not want a KG of artworks presented as facts, since they would not, and arguably should not, believe untraceable claims about a contested text. Their epistemic home is source work, the reading, citing, and criticizing of texts with direct reference to their origin. A representation that provides a source-anchored, criticizable trail aligns with this practice, whereas a fact-oriented graph would require abandoning the critical methods central to the discipline. Source criticism is not a constraint imposed by this dissertation; rather, it reflects the form of trust already practiced by the intended users.

This conceptual inversion underpins the four contributions discussed in Section 1.4: the PIEA, the provenance gap, hermeneutic provenance, and COMP-PROV. The methodological practices examined in the following section are derived from this foundation.

11.8 Can the digital humanities and the related challenges offer insights on best practices as solutions?

The unique challenges of DH projects (heterogeneous sources, interpretive complexity, evolving requirements, interdisciplinary stakeholders) make systematic engineering practices not optional but essential. The Viewsari project illustrates this through various observations, which we elaborate on below.

The Viewsari ontology underwent multiple design iterations (Section 7.3), each driven by new requirements that emerged primarily from domain-expert feedback. The initial formalization lacked the provenance layer entirely; the bibliographic layer was restructured when the FaBiO alignment revealed inadequacies in the original FRBR modeling; the content layer's mention typology was refined through the annotation campaign. The procedural framework that managed this evolution without losing coherence is the XD methodology of Presutti et al. (2009), Blomqvist et al. (2016), in which CQs serve as the primary input to each development iteration, and modular components of the ontology are released only once all associated CQs are answerable using representative test data. The dual role of CQs in that lifecycle, requirements in the upstream step and evaluation criteria in the downstream step (with the formal satisfaction condition that a SPARQL query against the populated graph returns a non-empty, correctly typed set of answers (Grüninger and Fox 1995, Blomqvist et al. 2016)), is exactly the role the 70 Viewsari CQs take in the present work. The post-closure CQ coverage reported in Section 9.1.3, with no question left unanswerable for modeling reasons, is the empirical signature that this lifecycle terminated honestly rather than by tolerance threshold, and the closure rules of Section 9.1.3, gap-recovery procedures expressed in the ontology's own vocabulary rather than additions to the TBox, are exactly the kind of artifact XD's iteration discipline is designed to produce.

DH projects are subject to frequent changes due to funding cycles, fluctuating team composition, and evolving RQs. The modular architecture of both the ontology (three independent layers) and the prompt templates is a direct response to this reality. When new entity types are added (as demonstrated by the Kloß pilot), when the underlying LLM changes, or when new evaluation criteria emerge, the changes are localized to specific modules rather than requiring a full system redesign. This modularity is not an architectural luxury but a sustainability requirement for projects that aim to outlast their initial funding period.

The epistemological commitments articulated in Chapter 2 are not preliminary considerations set aside during engineering; rather, they define the specifications that the engineering realizes. Principles such as source criticism, the distinction between statistical and epistemic uncertainty,

and the constructed nature of interpretive knowledge each impose concrete design requirements: a typed mention, a provenance activity, and an OOKB IRI. In this way, scholarly practice is encoded as a property of the data structure itself, not as an external guideline. A KG differs from a monograph precisely in this respect. It does not merely assert claims; it renders them open to direct inspection. This capacity is not an incidental technical feature but the enactment of an epistemological commitment. A reader who questions an attribution can re-execute the relevant query, examine the provenance activity, and amend or annotate the corresponding triple. The representation thus actively facilitates revision, whereas prose can only passively await it. Accordingly, the philosophy of science motivating this dissertation is realized through the artifact itself.

The contributions of Viewsari are most clearly articulated through the CQs. For art historians, tasks that previously required manual reading are now formalized as queries with explicit critical dimensions. For example, the question "Which artworks in this biography are named, and which only described?' (CQII.6) yields structured, inspectable results. Because the KG maintains visibility into the construction process, it enables further interrogation of evidentiary status, such as determining whether a relationship between two artists is grounded in explicit textual evidence, implicit inference, or statistical co-occurrence (CQII.40), and supports provenance tracing from entity annotation through to the original source (CQII.39). This level of source anchoring, required by art historical scholarship, is now realized explicitly. At a foundational level, Viewsari provides the first provenance-aware semantic representation of *The Lives*. At the same time, the extraction process is elevated to a first-class, queryable object, rather than remaining a byproduct. The system exposes the activity and software agent responsible for each entity, the prompt template and model version used, and the methodological origin – statistical or generative – of each extraction (CQII.34 to CQII.37). The ontology further distinguishes the epistemic weight assigned to PMI-based co-occurrences versus LLM-inferred implicit mentions (CQII.41).

Considered in a broader context, this architecture directly addresses the central methodological tension in computational literary studies: the opposition between close and distant reading introduced in the discussion of scalable reading (Chapter 1). Distant reading operationalizes a corpus as a quantifiable structure – manifesting as co-occurrence networks, character systems, or topic distributions – but does so at the expense of the situated interpretive act preserved by close reading (Moretti 2013, Jockers 2013, Jänicke et al. 2015). The principal critique from proponents of close reading is that such aggregation abstracts away from the textual grounding of interpretation. Viewsari's co-occurrence pattern directly addresses this critique by explicitly treating co-occurrences as observations about texts, not as events in the world, and by encoding, via COMP-PROV, the relevant span, the extraction run, and the responsible agent. This ensures that each distant-reading edge remains traceable to the specific close-reading judgment that

authorized it. Accordingly, Viewsari does not introduce another distant-reading technique; rather, it establishes a framework in which distant-reading artifacts remain accountable to the interpretive acts they summarize, extending the dissertation's methodological intervention from art history to computational literary studies.

11.9 Current limitations of the Viewsari framework

We complete this discussion with an assessment of the current limitations, each of which informs the future directions of work outlined in Section 12.3.

The empirical triples-per-mention ratio of approximately 9.7 (after closure) produces a graph that nonetheless grows substantially with corpus coverage. At the current (Population Run 2) scale of 86,654 mentions over the full 3,479-paragraph corpus, the populated graph lands at 2,455,976 triples after closure, and query performance over this graph size has not yet been benchmarked against the GraphDB endpoint under realistic concurrent access patterns; the materialized inference closure was therefore staged offline (Section 9.1.3) rather than at query time.

The three-layer architecture and the PROV-O activity chain, while justified by the provenance requirements discussed in Section 11.2, make the ontology non-trivial to query without domain knowledge. The SPARQL endpoint example queries (Section 7.7) partially address this; a more systematic guided query layer would improve accessibility for art historians without SPARQL expertise.

The current annotation and evaluation cover only the de Vere English translation, and the ObliquER pipeline is configured for artwork entities. Extension to the Italian original of the 1568 *Giuntina* edition, and to persons, locations, and motifs as primary extraction targets, requires a new GT annotation effort. The single-language scope also narrows the cross-version reading. The transferability discussion of Section 11.4 compares against ArtPedia (Section 9.4.7) as a period- and register-shifted English-language corpus, but cross-language and cross-translation comparison (e.g., de Vere English against the Italian original or against the Wagenbach German critical edition) would require a second domain-aligned GT and is not undertaken here. Both extensions are identified as the primary future work direction in Section 12.3. Furthermore, the construction of the ArtPedia dataset is based on the heuristic that the visual and contextual sentences are the sole such sentences in the paragraphs. However, this is not guaranteed, making false negatives possible.

The representational and methodological claims of this dissertation are evaluated against deliberately selected diagnostic cases rather than an exhaustive panel of vocabularies and corpora, because the contribution is representational and methodological rather than a benchmark. The

rationale for this design, the CIDOC CRM family as the best-case existing candidate so that the negative native-expressivity result on R1–R3 generalizes *a fortiori* to less expressive vocabularies (Section 11.2), and the Yin-style theoretical replication bracketed by a *near* case (ArtPedia) and a *far* case (the Kloß/FactGrid pilot) (Section 11.4), is set out in the research-question sections above and not repeated here. Broader cross-vocabulary alignment, additional target KBs, and a larger corpus panel are correspondingly future work (Section 12.3) rather than gaps in the present evidence.

The ObliquER pipeline emits explicit, implicit, coreferent, and generic artwork mentions and links each to a Wikidata QID or an OOKB IRI. It does not include an automated alignment step for complex iconographic references, motifs, or compositional descriptions. Passages that name a specific iconographic motif (e.g., a Last Supper composition or an Annunciation type) are currently solely recorded as the surrounding artwork mention and are not separately reconciled against a Wikidata motif entity, Iconclass (van de Waal 1974–1985) code, or other controlled vocabulary. Motif-level extraction would require both a dedicated annotation pass on the GT and an alignment stage that consults motif-specific authority resources; both are deferred to future work and bear on the persons, locations, and motifs extension named in the preceding paragraph.

The methodology (Section 7.4.3) envisaged a scalar per-mention confidence value $\kappa(m_j) \in \{\textit{high}, \textit{low}\}$ as output of the recognition stage, intended to record confidence on each typed mention. The current ontology and the populated KG do not realize this binding, and instead, we bind this epistemic qualification implicitly to the mention typology itself (Section 9.3), which records the manner of identification rather than a scalar self-report. The plan is to reinstate this value through possibly token-level log-probabilities or a calibrated post-hoc estimator that is not itself an LLM self-report (Kadavath et al. 2022, Xiong et al. 2024), and to evaluate against human-annotated confidence labels to verify that the values are interpretable (Piotrowski 2019, Schimmenti et al. 2026, Piotrowski 2026). Both are deferred to future work.

12 Summary

Knowledge extracted from interpretive texts is not discovered but constructed; therefore, KGs must model the construction process itself (Chapter 1). The main scientific contribution of this dissertation lies in an ontological and methodological framework that treats interpretations and their provenance as first-class citizens in the construction of a knowledge graph based on sources from the humanities.

Accordingly, we set out to answer this, a statement made implicitly in the dissertation's title. What does it mean to *model interpretation* when the interpreter is no longer exclusively human but increasingly a generative AI system? In the preceding chapters, working at the intersection of semantic technologies, DH, and GenAI, we have developed, instantiated, and evaluated a methodology that treats KE from historical sources not as fact retrieval but as a documentable, provenance-aware interpretive process. Giorgio Vasari's *Lives of the Most Eminent Painters, Sculptors, and Architects*, a text that has shaped art-historical interpretation for over 450 years, served as both the proving ground and the motivation for this work.

At the center is the distinction that conventional KG construction elides, the difference between a non-identity of a textual *mention* and the real-world *referent* it may or may not have (Chapter 1). We name this PIEA (C1). PIEA treats mention-to-referent resolution as completed before representation, materializing the referent directly from the mention and discarding both the textual occurrence and the act that recovered it. Viewsari's answer is structural, based on a three-layer architecture. We built it to sustain mention, referent, and bibliographic source as distinct first-class entities, with a four-way mention typology that records the kind of referent that was identified, and the provenance layer binds every mention to the paragraph, activity, model, and prompt template that produced it. Where a naive pipeline collapses the occurrence of "Pietà" into the sculpture (Section 1.2), Viewsari preserves the mention as a typed annotation whose referent is resolved, or left openly unresolved, under documented conditions: *ceci n'est pas une Pietà*, just as Magritte's painted pipe is not a pipe.

This commitment collects patterns in the KG that a fact-oriented representation may have discarded as noise. What becomes observable from this is that implicit reference and OOKB status track each other structurally. Entities that Vasari assumed his readers would recognize without naming are precisely those for which no canonical identifier was ever established, so that implicit artwork

mentions dominate the gold annotations, and roughly four-fifths of the resolved artwork entities in the populated graph are OOKB – up from about two-thirds in the gold subset (Section 11.1). A pipeline restricted to explicit mentions would therefore lose not a random fraction of the artwork population but the specific subset most tightly bound to the canonical narratives *The Lives* itself helped establish. This is not a data-quality problem but a *canonization signature* preserved in the KG, which this framework turns into a queryable property.

12.1 Revisiting the research questions

The following section revisits the three RQs posed in Section 1.3 to summarize the concrete contributions and assess the extent to which the design decisions – ontological, architectural, and methodological – hold up against the challenges identified.

RQ1: Representational commitments for interpretive accountability

The Viewsari ontology (Chapter 7, Chapter 8) provides a three-layer architecture (bibliographic, structural, content) that models the heterogeneity of Vasari's text at multiple levels of granularity, from the FaBiO representation of editions, translations, and volumes, through the DoCO-based (Peroni and Shotton 2018) segmentation of biographies and paragraphs, to the content layer that distinguishes explicit, implicit, coreferent, and generic artwork mentions.

Heterogeneity in historical sources is fundamentally epistemic as well as structural. The same artwork may appear canonically named in one passage and only obliquely described in another; attributions may be asserted as fact or qualified as opinion. The Viewsari ontology addresses this epistemic variability by integrating PROV-O (Lebo et al. 2013) and the Web Annotation Ontology (Sanderson et al. 2017) as core components rather than some metadata. Each extracted triple is accompanied by provenance metadata that links it to the source paragraph, the extraction model, and the prompt template version. The typed mention class records the identification method. Comparative modeling against CIDOC CRM (Doerr 2003) (Section 9.2) demonstrates that this integration is essential: CIDOC CRM cannot formally represent implicit entity status, extraction provenance, or epistemic qualification. The 100 % provenance coverage in the populated KG (Section 9.5.4) establishes this as an empirical property of the system, not merely a theoretical assertion.

The ontology was developed using the XD methodology (Presutti et al. 2009, Blomqvist et al. 2016) in close collaboration with domain experts, ensuring that the formal model is anchored

in art-historical research practice. The 70 competency questions (Chapter A) and their coverage rates with respect to the populated KG serve as a quantitative measure of this alignment.

These representational commitments render within-corpus epistemic heterogeneity explicit at the data level. The per-biography decomposition of recognition behavior (Section 9.4.3) can be recovered from the populated graph via the aligned bibliographic chain and mention-level provenance edges. As a result, the corpus heterogeneity identified in Chapter 2 as constitutive of humanities sources is preserved and remains queryable across biographies, rather than being collapsed. The Michelangelo outlier discussed under RQ2 (Section 11.3) exemplifies this. The most canonized artist in *The Lives* is associated with the densest body of descriptive, implicitly referring prose and is consequently the most challenging to extract. Canonization thus increases, rather than reduces, recognition difficulty, and the per-biography decomposition provides precisely the diagnostic that the ontology was designed to enable.

RQ2: Large language models as interpretive agents for implicit and out-of-knowledge-base entities

With ObliquER (Chapter 7, Section 7.4 to 7.4.9), we demonstrate that generative extraction extends to art-historical narrative sources in ways that surface form-level classifiers cannot. The principal contribution is the recognition and linking of implicit mentions of artworks. We built a system that can extract and link especially descriptive phrases such as "a panel for the friars of Santa Croce, representing the Last Supper", which refer to specific, recoverable artworks without explicit naming. Surface form-level NER systems such as UniNER (Zhou et al. 2024c) and GLiNER2 (Zaratiana et al. 2025) achieve 0% recall on this class by design (Section 9.4.2). Their scores of 0.00 to 0.02 do not represent a threshold that ObliquER surpasses, but rather indicate the extent to which descriptive reference lies outside the reach of surface form recognition. ObliquER's Strict F1 of 0.321 is therefore not a peak benchmark result, but the first non-zero measurement on this recognition surface. The substantive contribution is the modeling that renders this surface addressable and accountable, rather than a particular performance.

With the prompt engineering methodology, we added another layer of modularity to the framework; the template architecture separates domain-specific content, such as entity definitions, special cases, and few-shot examples, from structural components. This separation enables systematic ablation and domain transfer without requiring changes to the pipeline. The five-block prompt design demonstrably improves extraction quality over an unstructured LLM baseline (Section 9.4.6). The cross-model evaluation (Section 9.4.5) extends the prompt-architecture hypothesis from a single-decoder context to a cross-decoder context. Under both `google/gemma-4-31b-it` and `openai/gpt-oss-120b`, the recognition surface remains stable, an invariance that aligns

with the canon-convergence thesis of Chapter 2, where instruction-tuned decoders derived from a shared foundation-model lineage are expected to cluster. With Viewsari, rather than flattening a claim's generative conditions, we explicitly make these conditions a part of the main KG, e.g., with the typed `prov:SoftwareAgent`.

We designed the entity linking pipeline to address a corpus characterized by the long tail of historical artworks, with about two-thirds of artwork entities absent from Wikidata (Vrandečić and Krötzsch 2014) in the gold GT subset and roughly four-fifths (79.3 %) in the populated graph. This distribution constitutes the 'sea of unknowns' described in Section 7.4.9. Our system does not treat OOKB status as a failure. Instead, it assigns stable Viewsari-internal IRIs to these entities as first-class KG nodes, making them available for future reconciliation. This is achieved through a retrieve-then-rank strategy, in which complementary dense and surface form retrievers provide candidates to an LLM-based ranker that outputs either a Wikidata identifier or an OOKB verdict.

ObliquER's outputs across all ten volumes, combined with the manually annotated ground truth (16 biographies, 232 paragraphs) and the foundation layer, populate a KG comprising 2,455,976 triples after closure. This includes 86,654 typed mention annotations, 15,804 artwork entities, 443 persons, and 632 co-occurrences, all under complete provenance coverage. These figures do not serve as benchmarks, but rather as evidence that the methodology scales to a substantial corpus while maintaining its epistemic commitments.

RQ3: Design patterns and methodological transferability

We assessed the transferability of the Viewsari methodology along a comparative analysis of ODPs across DH KGs and a pilot study that reapplies the pipeline to a structurally distinct corpus (Chapter 10).

The comparative pattern analysis found that across Viewsari and three comparison projects (ICON (Sartini et al. 2023), Polifonia (de Berardinis et al. 2023), Odeuropa (Lisena et al. 2022)) only the Provenance pattern is genuinely shared, confirming that DH ontologies remain substantially domain-tailored. The result of this detailed analysis, COMP-PROV, understood as a specialization of the Annotation pattern that binds its *agency* dimension to a parametrized software agent, captures the machine-generated nature of extraction that the comparison projects' manual or rule-based processes do not raise. Read in continuity with Ondraszek et al. (2024c), we contribute the ANN pattern, COMP-PROV, and the Co-occurrence content pattern (Section 10.3.3), which types textual co-occurrences as `prov:Entity` rather than as `Event` specializations. With the Annotation pattern, the three form a coherent bundle for DH KGs whose content consists of annotations on extracted data: ANN anchors a statement to its source span, Co-occurrence types

the observation as an artifact of a text rather than an event in the world, and COMP-PROV records the activity-level conditions under which the annotation was produced.

The *Kloß Annalen pilot study* (Section 10.2.1) tested the ObliquER prompt architecture on a fundamentally different source: 19th-century German-language documentary prose on the Illuminati and Freemasonry, linked to FactGrid (Simons 2024) rather than Wikidata. The adaptation required changes exclusively to the static content layer (role descriptions, entity definitions, and few-shot examples), while the structural templates, the inference pipeline, and the post-processing logic remained unchanged. The adapted schema extended the Viewsari mention typology with domain-specific types (coded aliases, coded locations) that map structurally onto implicit mentions: in both, the surface form does not transparently identify the referent, and domain knowledge is required for resolution.

Accordingly, the 64-paragraph Kloß GT (237 annotated mentions across persons, organizations, and locations) shows that the modular architecture generalizes across language, period, entity type, KB, and referential strategy. The pilot also instantiates a two-tier linking design (FactGrid $\rightarrow$ OOKB) that swaps the first-tier KB for a domain-specialized alternative, but its end-to-end evaluation is reserved for future work (Section 11.9): the pilot demonstrates recognition transfer and serialization into the FactGrid-aligned schema, not a measured out-of-domain OOKB rate. The design intent nonetheless runs counter to the 79.3 % OOKB artwork-entity rate on Vasari relative to Wikidata. Even a deliberately long-tail historical KB with no notability bar is not expected to close the coverage gap, so a KG architecture that treats OOKB entities as first-class outputs is the structurally correct response in both the general-purpose and purpose-built regimes.

The transferability claim is also quantitatively visible on ArtPedia (Section 9.4.7). Under the same production prompt tuned against Viewsari, ObliquER reaches Strict F1 $= 0.319$ against Viewsari's 0.334, a 1.5-percentage-point drop across a corpus shift spanning period, register, and scale, so the ontology-grounded recognition surface that distinguishes ObliquER from the surface form-only baselines transfers cleanly. On the EL side, the regime distinction sharpens (Table 9.12): explicit-mention linking is effectively saturated across all four systems, while in the implicit subset, ObliquER's advantage is preserved and amplified, separating it from the next-best baseline by an order of magnitude.

Composition of the contributions

Table 12.1 consolidates our mapping between RQs and contributions and their interplay.

E1, the Viewsari ontology, encodes C1's rejection of the PIEA. Our evaluation demonstrates that this encoding is logically consistent. The TBox, verified consistent by HermiT, maintains the

Table 12.1: Each research question is answered jointly by a conceptual commitment (C1–C4) and the empirical artifact (E1–E4) that operationalizes it. The table consolidates the principal evidence; the composition is read as a single argument in Section 12.1.

RQ	C	E	Principal evidence
RQ1	C1	E1	100 % provenance coverage; zero unanswerable CQs of 70; HermiT-consistent OWL 2 DL TBox; CIDOC CRM (with PROV-O) and CRMinf cannot express span anchoring, extraction provenance, and OOKB status simultaneously
RQ2	C2, C3	E2, E3	Recognition capability Partial / Type F1 0.697 / 0.642; implicit Strict F1 0.321 against 0.00–0.02 for surface form-level baselines; implicit KB-linked F1 0.385 against ≤ 0.06; ontology-typed prompt carries 72–95 % of the content-layer gain; recognition stable across decoders (0.048 envelope)
RQ3	C4	E4	COMP-PROV unique among four DH KGs (only PROV shared); Kloß recognition span-based Strict F1 0.600 / 0.683; ArtPedia Strict F1 0.319 against Viewsari's 0.334 under the same prompt

mention/referent separation without introducing any empty classes. The 70-question competency question catalog yields no unanswerable cases. Furthermore, the comparative modeling study confirms that neither CIDOC CRM, CIDOC CRM combined with PROV-O, nor CRMinf can simultaneously represent span anchoring, computational-extraction provenance, and OOKB status, as detailed in Section 9.2. E2 and E3 operationalize C2 and C3, which address the provenance gap and its fundamentally hermeneutic, rather than mechanistic, nature. The resulting graph achieves complete provenance coverage. The strict M07 ratio of 0.751 does not indicate a deficiency; rather, it reflects a graph that intentionally encodes complex extraction details. This partition is interpretable only because C4's COMP-PROV pattern reifies the extraction agent as a typed agent node. The recognition metrics reinforce this interpretation. ObliquER's implicit Strict F1 of 0.321, compared to 0.00 to 0.02 for surface form-level baselines, does not represent a competitive ranking; instead, it empirically establishes the recognition surface that C1 posited. When the boundary-tolerant Partial and Type criteria are applied, the surface measures 0.697 and 0.642, respectively, once the exact-span penalty – incurred even by human annotators (0.43 versus 0.83 partial) – is excluded. Component ablation further attributes this improvement to the modeling commitment itself, rather than to prompt optimization.

C4 and E4 complete the argument by demonstrating that this apparatus constitutes a reusable methodology rather than a Vasari-specific artifact. The comparative pattern analysis in Section 10.3 identifies COMP-PROV as unique among four digital humanities KGs, with the Annotation pattern's agency dimension assigned to a parametrized software agent, whereas previous projects consistently assigned this role to a person. In the Kloß pilot, the ontology core, prompt templates, post-processing, and provenance overlay remain fixed; only the prompt content layer is modified. Both components transfer in tandem. Recognition performance is maintained, and the ontology accommodates new mention types. Considered together, the four contributions are

not merely concurrent; each empirical artifact embodies a specific conceptual commitment, and the evaluation metrics ensure alignment between the conceptual and artifactual dimensions of the dissertation.

12.2 Summary of contributions

Stated as one claim, we define the core contribution of this dissertation as an extension of (historical) source criticism to machine extraction, delivering the ontology (Viewsari) and the design pattern (COMP-PROV) that let a KG treat each claim a generative model extracts as a source to be criticized rather than a fact to be trusted.

Returning to the title of this dissertation, *Modeling interpretation in the age of GenAI*, we summarize this contribution across modeling, interpretation, and an AI perspective.

12.2.1 Modeling

As we have stated right at the beginning in Chapter 1, the goal of our work is conceptual rather than that of a typical NLP benchmark work. Thus, the Viewsari ontology provides a formally defined, provenance-aware knowledge representation that distinguishes structurally grounded triples (bibliographic metadata) from interpretively grounded triples (extracted entities with documented epistemic conditions). As we have shown through the architectural ablation (Section 9.4.6), the ontology-typed prompt structure, not the demonstration budget or the choice of decoder, carries the recognition gain over surface form-level-only baselines. Our Viewsari modeling commitment is what gets the work done, so to speak.

The comparative modeling study (Section 9.2) situates this contribution against the strongest available alternative, the CIDOC CRM family. The combinatorial objection that such a stack might recover the same representational ground as Viewsari is addressed directly in Section 9.2.2 and shown to rest on a layer mismatch rather than a feature gap. That same layer difference is itself the positive contribution and not merely a rebuttal, since Viewsari occupies the recognition-and-linking layer beneath the cultural-heritage stack, on whose resolved referents a CIDOC event layer can build (Section 11.5). CIDOC CRM's existential presupposition, that every modeled entity exists or has existed in the world, is structurally unavailable for the thousands of implicit OOKB mentions whose referents are not independently attested, requiring the mention typology, too.

12.2.2 Interpretation

ObliquER operationalizes the thesis that extraction *is* interpretation. By recording the prompt template hash, the model identifier, and the typed mention class, the full reasoning trace on each `prov:Activity` node, the KG documents not only *what* was extracted and *from where*, but *why* the model concluded that a given phrase constitutes an artwork reference. The interpretive act of the generative model is thereby made formally equivalent to a human annotator's decision, since both are documented `prov:Activity` instances with identified agents, recorded inputs, and traceable outputs.

12.2.3 The age of GenAI.

The dissertation engages directly with the epistemic consequences of using LLMs for knowledge construction in the humanities. It demonstrates that the outputs of generative systems are not facts to be accepted or errors to be filtered, but probabilistic interpretive acts whose conditions of production must be modeled as constitutive components of the resulting KG – similarly to how it has been done in the more traditional approach of source criticism. The cost we pay for the provenance gap is accountability. Undocumented, a model's hallucination or, when returning to historical sources and their interpretation, a scholar's attribution enter the record indistinguishably, and claims become facts. *Extraction without documentation is epistemology without accountability.* We deliver these concrete engineering responses in the form of the COMP-PROV ontology pattern, the prompt-inclusive provenance chain, and the OOKB entity handling. They ensure that the KG remains transparent and contestable, properties that Chapter 1 identified as prerequisites for scholarly use.

The CQ catalog (Section A) makes these capabilities explicit. For example, an art historian can query the graph to determine which works in a biography are named and which are only described (CQII.6). The historian can then trace any answer through annotation, paragraph, edition, and translation to the original source (CQII.39), or assess whether a claimed relationship is based on explicit text, implicit inference, or mere co-occurrence (CQII.40), thereby transforming source-critical reading into a query. Similarly, a knowledge engineer can identify which activity, prompt template, and model version produced a given entity (CQII.34, CQII.35, CQII.37) and examine the extraction process as data.

12.3 Future work

For future work, we deem different extensions necessary. The most immediate is multilingual and multi-entity coverage. We plan to extend ObliquER to the Italian original of *The Lives* and to additional entity types, including persons, locations, and iconographic motifs. The modular prompt architecture is designed to support such extensions by incorporating new entity definitions and examples into the configuration, without requiring changes to the template or pipeline. The Kloß pilot has already demonstrated this approach for persons, organizations, and locations in German.

Relation extraction is another such extension, repeatedly requested by domain experts, for capturing more semantically meaningful connections between entities that, so far, are only established via co-occurrence or shared appearance in the same paragraphs or at higher levels of the document/bibliographic structure.

A further methodological refinement concerns linking precision at both the candidate-generation and ranking stages. Our entity-linking error analysis (Section 9.4.4) surfaced over-predictions in which the ranker selected a candidate whose date and place lie outside the world of *The Lives*; in one case, a Vasari mention was linked to `Q125760953` (*Fuente del Triunfo*), a fountain created in Córdoba in 1765. Such candidates are recognizable as impossible on the metadata the pipeline already retrieves, because *The Lives* documents artworks constrained by Vasari's lifetime (terminating in 1574) and by the Italian locales he described. The structured candidate records include inception (`wdt:P571`) and location (`wdt:P276`) fields (Table 7.13) that currently inform the ranker only in a soft descriptive capacity. Imposing them as hard spatiotemporal constraints, namely a creation date not later than Vasari's death and a geographic prior aligned with the corpus, would prune anachronistic and geographically implausible candidates prior to ranking. This approach provides the LLM ranker with an explicit admissibility criterion, rather than relying on implicit inference. Crucially, these constraints are defined at the corpus level and are not Vasari-specific heuristics. Reparametrizing the temporal and geographic bounds enables the same mechanism to generalize to other corpora, as demonstrated in the Kloß pilot, where the relevant envelope is defined by the late eighteenth-century Illuminati order and German geography.

A lower-priority extension concerns the bibliographic layer. Since the bibliographic hierarchy underlying Viewsari is preserved in the IFLA LRM, migrating the FaBiO/FRBR layer to an LRM-based vocabulary would require only namespace-prefix changes rather than a structural redesign. We therefore identify this as an optional future alignment, to be pursued if full library-catalog interoperability becomes necessary.

Furthermore, we aim to focus on using the populated graph for content-focused analyses, i.e., entity co-occurrence. This structure enables a new analytic layer that operates independently of the extraction process, including network analysis, statistical and distributional studies (Amancio 2016, Jost and Jost 2019) of the canonization signatures discussed in Section 11.1, and the development of a confidence calculus that aggregates per-activity evidence into derived estimates. Extending this analytic layer to encompass not only person co-occurrences but also artworks and their iconographic relations is a logical next step. The COMP-PROV provenance trail is specifically designed to ensure accountability, as any derived result can be traced back to the extractions it summarizes. This could also be taken up through reasoning layers. The Viewsari ontology is reasoning-ready, although we, in the present work, do not apply entailment at scale (Section 11.2). A profile-restricted, materialization-based regime, such as OWL 2 RL (Motik et al. 2012), exemplified by the competency-question closure procedure, would enable inference over the full graph when description-logic classification is not feasible (Urbani et al. 2009). This approach would support reasoning directly on the extracted claims, including detecting contradictions across biographies, editions, and translations; inferring artwork and actor types that the extractor leaves implicit; and propagating coreference and `owl:sameAs` links that remain unmerged after the linking stage. Each inferred triple constitutes an interpretive act and would be recorded as a distinct COMP-PROV activity, ensuring that derived knowledge remains as accountable as extracted knowledge.

12.4 A panel without a name: closing reflections on interpretive accountability

In Chapter 1, we established that digitization alone does not guarantee interpretability and that knowledge derived from historical texts is constructed rather than discovered. Throughout this dissertation, we have systematically pursued the methodological and engineering implications of this foundational observation.

When Vasari wrote that Cimabue painted “a panel for the friars of Vallombrosa, in the Abbey of S. Trinita in Florence”, he described an artwork without naming it, because naming things had not been en vogue when Cimabue practiced, and Vasari knew that true art connoisseurs would know what he described. This implicit reference that four and a half centuries of art-historical scholarship have learned to interpret through contextual reasoning, comparative analysis, and accumulated domain knowledge. Generally, historical descriptions should not be linked directly to real-word objects, but rather through explicitly modeled acts of interpretation. Viewsari implements this principle by separating the mention, the interpretation, and the referent. When ObliquER identifies this phrase as an implicit artwork mention, links it to a Wikidata entity, and

records the extraction as a documented `prov:Activity` with a typed mention class, a model identifier, and a prompt template hash, it performs a structurally analogous interpretive act whose conditions of production are now transparent and reproducible.

The Viewsari methodology does not claim that this computational interpretation is equivalent to a scholar's – we already make enough other daring claims in this dissertation. What we do claim, however, is that in the age of generative AI, where LLMs increasingly participate in constructing knowledge from humanistic sources, the conditions under which these interpretations are produced must be modeled with the same rigor we demand of the knowledge itself. Semantic technologies provide the formal apparatus for this modeling. The contribution of this dissertation is to have shown, concretely and at scale, that such modeling is not only desirable but feasible. Interpretation can be made a first-class object in knowledge engineering, yielding KGs that are richer, more transparent, and more epistemically honest. Seen in the long view, this is the discipline of historical source criticism extended to the machine; the same scrutiny historians have applied to a source, documenting conditions, actors, and purpose, is here rendered computable and applied to the generative model, so that its extractions enter the scholarly record as sources to be criticized rather than as facts to be trusted. The construction of a KG from sources in the humanities must not obscure interpretation; rather, it must make interpretation itself the subject of the representation. Thus, the provenance of a reference statement is a standalone object of knowledge and not merely metadata.

Modeling interpretation in the age of GenAI is, ultimately, exactly this acknowledgment. Our work foregrounds the explicit modeling of interpretations as independent ontological objects within the KG framework. Whether we manually analyze sources or extract entities from them with a computational pipeline, we engage in an interpretive act. Whether performed by a Renaissance biographer with a thing for Michelangelo, an art historian, or a twenty-first-century LLM, this act deserves to be documented as such. Whatever enters the scholarly record must be visibly contestable, not disguised as a fact. The panel Cimabue painted for the friars of Vallombrosa has no canonical name, and perhaps it never will. But the conditions under which it was recognized, described, and entered into a KG are now on record. That is what we owe our past and future cultural heritage, not the pretense of certainty, but an anchor – a documented, contestable claim – so that what we cannot name, we can at least not lose.

A Appendix

In this appendix, we present the complete, structured requirements catalog for the Viewsari ontology, systematically derived from two phases of requirements engineering in collaboration with art historical domain experts. The Phase I requirements (`RQI`) define the foundational scope established during the initial collaborative sessions. The Phase II requirements (`RQII`) reflect the outcomes of iterative expert consultations as we extended the project to address large language model-guided entity extraction.

Phase I requirements (`RQI`)

Phase I requirements reflect the initial elicitation round, grounded in the social network and prototype knowledge graph produced by Santini et al., and structured into two groups: data and provenance requirements, and user-facing functionality requirements.

Data and provenance

ID	Requirement
`RQI.1`	Representation of persons (artists) and all other extracted entity types—including organizations, places, artworks, motifs, terms, and dates—as structured, individually addressable resources.
`RQI.2`	Typed relations between all entity types, including persons, artworks, and locations (e.g., mentorship, co-occurrence).
`RQI.3`	Provenance and positional information for all extracted entities and their relations, connected to the source paragraph, biography, and top-level bibliographic metadata of the work (formal representation of Vasari's original text, distinguished from its various editions and translations).

ID	Requirement
RQI.4	Direct links to original text passages from *The Lives*, including the Italian source text, and to external authority data such as Wikidata.
RQI.5	Representation of relevant attributes of resources: birth and death dates for persons; creation dates for artworks.
RQI.6	Representation of historical events and their temporal context, connected to relevant persons, artworks, and locations.
RQI.7	Support for fact-checking and federated queries across external datasets such as Wikidata.

User-facing functionality

ID	Requirement
RQI.8	Index searches for extracted resources: artists, artworks, locations, etc.
RQI.9	Support for semantic and visual search.
RQI.10	Exploration of resources and co-occurrences without a pre-formed research question (open-ended browsing).
RQI.11	Visibility of contextual implications and the connection strength between resources.
RQI.12	Support for interpretation and discovery through the visualization of resources and their connectedness.

Phase II requirements (RQII)

Phase II requirements extend the initial elicitation to reflect the expanded scope of the project, in particular the modeling of large language model-guided generative extraction and provenance. They are organized into eight thematic clusters identified collaboratively with domain experts using the OntoChat framework.

Cluster 1 – Artists, collaborations, and relationships

ID	Requirement
RQII.1	Representation of various types of associations between artists, distinguishing statistical co-occurrence from typed (explicit) relationships.
RQII.2	Storage and queryability of quantitative co-occurrence indicators, including pointwise mutual information (PMI) scores, Dice coefficients, and network centrality values.
RQII.3	Representation of artist metadata: canonical names, alternative surface forms, birth and death dates, and occupations.
RQII.4	Distinction between the `viewsari:Artist` subclass and the broader `viewsari:Person` class, enabling queries that differentiate co-occurrences involving two artists from those involving an artist and a non-artist (e.g., a church official or patron).

Cluster 2 – Artworks

ID	Requirement
RQII.5	Dedicated representation of artworks with attributes including title, creation date, genre, and medium.
RQII.6	Ability to explore artworks in relation to their creators and their depictions in Vasari's text.
RQII.7	Identification and representation of relationships between identical or different entities across artwork instances (e.g., shared iconographic motifs, co-location, or stylistic affinity).
RQII.8	Representation of both explicit and implicit artwork mentions, with each mention anchored to the source text span and classified by mention type (*explicit*, *implicit*, *coreferential*).

Cluster 3 – Locations

ID	Requirement
RQII.9	Representation of locations mentioned in *The Lives*, including places where artworks were created or placed, and where artists lived and worked.
RQII.10	Support for exploration and retrieval of locations and their textual depictions in Vasari's text.
RQII.11	Linkage of location entities to external authority sources (e.g., Wikidata) for geographic disambiguation and federated queries.

Cluster 4 – Historical events

ID	Requirement
RQII.12	Representation of historical context, including major events (e.g., the Catholic Reformation, the invention of the printing press, the birth of a contemporary artist).
RQII.13	Representation of transaction-type events such as commissions and the destruction or relocation of artworks, with temporal attributes.
RQII.14	Integration of historical event information from external sources for context enrichment.

Cluster 5 – Evidence and bibliographic information

ID	Requirement
RQII.15	Distinction between work-level, expression-level, and manifestation-level bibliographic metadata, aligned with the FaBiO/FRBR four-layer model (*Work*, *Expression*, *Manifestation*, *Item*).
RQII.16	Representation of dependencies between extracted content and its source, making explicit that all extraction is performed on de Vere's 1912 English translation rather than the original Italian edition.

ID	Requirement
RQII.17	Direct, clickable links to the exact web page hosting the relevant paragraph in the Project Gutenberg online edition (domain-expert requirement: bibliographic representation at the web-page level).
RQII.18	Accommodation of multiple concurrent editions and translations, enabling comparative work across versions of *The Lives*.
RQII.19	Representation of the provenance chain of the extraction process itself: which model or script produced the extraction, in which run, and through which prompt formulation—grounded in PROV-O.

Cluster 6 – Linked data enrichment

ID	Requirement
RQII.20	Enrichment of entities from *The Lives* with additional information from Wikidata, Iconclass, or similar authority sources.
RQII.21	Support for fact-checking against external knowledge bases.
RQII.22	Support for federated SPARQL queries across linked datasets.
RQII.23	Preservation of out-of-knowledge-base (OOKB) entities as provisional nodes with local identifiers, avoiding forced normalization.

Cluster 7 – User interaction and exploration

ID	Requirement
RQII.24	Support for directed (index) search for artists, artworks, and locations.
RQII.25	Support for open-ended visual exploration of the network without a pre-formed research question.
RQII.26	User interface accessible to users with and without prior domain knowledge.

ID	Requirement
RQII.27	Compliance with FAIR data principles, enabling researchers to reuse data in their own projects.
RQII.28	Availability of a SPARQL endpoint and/or RDF dump for programmatic data access.

Cluster 8 – Visualization

ID	Requirement
RQII.29	Graph-based visualization of co-occurrences and typed relationships.
RQII.30	Ability to explore potential biases and patterns in the data visually (e.g., distributional skew toward canonical figures).
RQII.31	Views illustrating instances where artworks reflect collaborative effects (e.g., shared locations, artistic styles, or workshop affiliations).
RQII.32	Interactive visualizations suitable for use in lectures and presentations, communicating complex relational and historical patterns to students.

Competency question catalog

This appendix provides the full catalog of competency questions (CQs) for the Viewsari ontology, derived from user stories collected over two requirements-engineering phases. Each CQ is assigned a unique identifier and, where applicable, linked to the thematic cluster and persona from which it originated. CQs are the formal operationalization of user stories and serve as both the basis for ontology design decisions and the evaluation instrument for assessing coverage in Section 9.5.

Phase I CQs (`CQI.1-29`) were derived from 32 user stories across four personas, which yielded 39 formally elicited questions; ten of these are functional web-application questions documented with the system layer (Section 7.7), leaving 29 catalog entries. Phase II CQs (`CQII.1-41`) add 15 further user stories, elicited with OntoChat and consolidated from 74 raw questions into 41 catalog entries, and address generative extraction provenance, bibliographic depth, and epistemic modeling. The catalog therefore comprises **70 competency questions** in total (29 Phase I + 41 Phase II), of which 52 are SPARQL-answerable (`select`/`ask`) and 18 are descriptive.

Type: the **Type** column records how each CQ is answered. `select` and `ask` denote SPARQL-answerable questions (52 in total), backed by an executable query in Section A; `descriptive` denotes functional questions (18 in total) – concerning data reuse, teaching, visualization, infrastructure, or modeling rationale – that are answered by documentation, tooling, or qualitative argument rather than a graph pattern, and therefore carry a short rationale in place of a query.

Personas: Dr. Elena Rossi (art historian researcher), Prof. Nazeera Marfi (professor of digital art history), Aaron Warner (computer scientist / knowledge engineer), John Saffron (art history student).

Phase I competency questions (`CQI`)

Artists and co-occurrences

ID	Competency question	Type	Persona
`CQI.1`	Who co-occurred with whom in *The Lives*?	`select`	Elena Rossi
`CQI.2`	Did artist X co-occur with artist Y?	`ask`	Elena Rossi
`CQI.3`	Which artists are mentioned together in the same passage?	`select`	Elena Rossi
`CQI.4`	Where in the text did artists X and Y co-occur? (In which paragraph, biography, volume?)	`select`	Elena Rossi
`CQI.5`	How frequently do two given artists co-occur across the entire corpus?	`select`	Elena Rossi
`CQI.6`	Which artists have the highest co-occurrence frequency with artist X?	`select`	Elena Rossi
`CQI.7`	What is the pointwise mutual information (PMI) score for the co-occurrence of artists X and Y?	`select`	Aaron Warner
`CQI.8`	What is the network centrality of artist X in the co-occurrence network?	`select`	Aaron Warner

Persons and attributes

ID	Competency question	Type	Persona
`CQI.9`	What are the birth and death dates of person X?	`select`	Elena Rossi
`CQI.10`	Which persons mentioned in *The Lives* are identified as artists?	`select`	Elena Rossi
`CQI.11`	Which persons mentioned in *The Lives* are not artists (e.g., patrons, clergy, scholars)?	`select`	Elena Rossi
`CQI.12`	In which biography (biographies) does person X appear?	`select`	John Saffron
`CQI.13`	Is person X linked to an external authority record (e.g., Wikidata)?	`ask`	Aaron Warner

Artworks, locations, and entities

ID	Competency question	Type	Persona
`CQI.14`	Which artworks are mentioned in the biography of artist X?	`select`	Elena Rossi
`CQI.15`	Which locations are associated with artist X?	`select`	Elena Rossi
`CQI.16`	Which artworks were created in location L?	`select`	Elena Rossi
`CQI.17`	Can students explore artist collaborations on a geographical map?	`select`	Nazeera Marfi
`CQI.18`	Where is location X located on a map?	`select`	John Saffron

Bibliographic and provenance

ID	Competency question	Type	Persona
`CQI.19`	Which paragraph and page in *The Lives* is the source of co-occurrence C?	`select`	Elena Rossi
`CQI.20`	Which volume of *The Lives* contains the biography of artist X?	`select`	John Saffron
`CQI.21`	What edition or translation of *The Lives* was used for extracting a given co-occurrence?	`select`	Elena Rossi
`CQI.22`	What is the original text snippet associated with co-occurrence C?	`select`	Elena Rossi
`CQI.23`	What licence is provided by the Viewsari dataset and where can the original source text be found?	`select`	Aaron Warner

Search and data reuse

ID	Competency question	Type	Persona
CQI.24	How can I execute semantic searches over the knowledge graph?	descriptive	Aaron Warner
CQI.25	How can I reuse the Viewsari data in my own project, and in what form can it be shared?	descriptive	Nazeera Marfi
CQI.26	Is there a SPARQL endpoint or an RDF dump available?	descriptive	Aaron Warner
CQI.27	Which external resources (ontologies, authority data) were used in constructing the knowledge graph?	select	Aaron Warner
CQI.28	Which authority data (e.g., Wikidata, Iconclass) is connected to the resources?	select	Elena Rossi
CQI.29	How can my students reuse data in the project course, and can the visualizations be modified and customized?	descriptive	Nazeera Marfi

Phase II competency questions (CQII)

Phase II CQs address the extended requirements that emerged as the scope expanded to include large language model–guided extraction, bibliographic depth, epistemic provenance, and out-of-knowledge-base entity handling. They are organized into the eight thematic clusters of the Phase II requirements.

Cluster 1 – Artists, collaborations, and relationships

ID	Competency question	Type	Persona
CQII.1	What is the difference between a statistical co-occurrence and an explicitly typed relationship between two artists in the knowledge graph?	ask	Aaron Warner

ID	Competency question	Type	Persona
CQII.2	Which association types between artists – statistical co-occurrence versus explicitly typed relationship – are represented in the knowledge graph, and how many instances does each carry?	select	Aaron Warner
CQII.3	Which artists have both a statistical co-occurrence and an explicit typed relationship with artist X?	select	Elena Rossi
CQII.4	What is the Dice coefficient for the co-occurrence of artists X and Y, and how does it compare to the PMI score?	select	Aaron Warner
CQII.5	How do I explore relationships between Renaissance artists, artworks, historical events, or locations so that I can identify patterns and trends for research or in a course?	descriptive	Nazeera Marfi

Cluster 2 – Artworks

ID	Competency question	Type	Persona
CQII.6	Which artworks are explicitly named in the biography of artist X, and which are described only implicitly?	select	Elena Rossi
CQII.7	What is the surface form by which artwork A is referred to in paragraph P, and is it an explicit, implicit, or coreferential mention?	select	Elena Rossi
CQII.8	Which artworks in the knowledge graph are classified as out-of-knowledge-base (OOKB) entities—i.e., have no match in Wikidata?	select	Aaron Warner
CQII.9	Which artworks co-occur with a specific artist in a particular volume, and what are the source paragraphs?	select	Elena Rossi
CQII.10	Can I explore how artworks reflect collaborative effects, such as shared locations or artistic styles?	descriptive	Nazeera Marfi

Cluster 3 – Locations

ID	Competency question	Type	Persona
`CQII.11`	Which locations are mentioned in co-occurrences involving artist X?	`select`	Elena Rossi
`CQII.12`	Which artists were active in location L according to *The Lives*?	`select`	Elena Rossi
`CQII.13`	Is location L linked to an external authority record (e.g., Wikidata, GeoNames)?	`ask`	Aaron Warner
`CQII.14`	Can students explore artist collaborations on a geographical map, filtered by location?	`descriptive`	Nazeera Marfi

Cluster 4 – Historical events

ID	Competency question	Type	Persona
`CQII.15`	Which historical events are associated with artist X in the knowledge graph?	`select`	Elena Rossi
`CQII.16`	Which artworks were commissioned or destroyed as part of a specific historical event E?	`select`	Elena Rossi
`CQII.17`	How can I identify cross-disciplinary connections between Renaissance art and other historical or cultural phenomena using the knowledge graph?	`descriptive`	Nazeera Marfi
`CQII.18`	Can information about historical events be enriched from external sources such as Wikidata, and if so, which properties are aligned?	`descriptive`	Aaron Warner

Cluster 5 – Evidence and bibliographic information

ID	Competency question	Type	Persona
CQII.19	At which bibliographic level (work, expression, or manifestation) is each entity of a given source recorded, for citation and attribution across editions and translations?	select	Elena Rossi
CQII.20	Which edition and translation of *The Lives* is the source of a given extracted entity or co-occurrence?	select	Elena Rossi
CQII.21	What is the direct web URL to the paragraph in the Project Gutenberg edition from which entity E or co-occurrence C was extracted?	select	Elena Rossi
CQII.22	Is the extracted information derived from de Vere's 1912 English translation, the 1568 Italian edition, or another version?	select	Elena Rossi
CQII.23	How does the choice of translation (de Vere's English vs. the original Italian) affect the entities and co-occurrences that appear in the knowledge graph?	descriptive	Elena Rossi
CQII.24	For co-occurrences involving artist X, can the English translation's web manifestation be compared with the Italian edition's web manifestation?	select	Elena Rossi

Cluster 6 – Linked data enrichment

ID	Competency question	Type	Persona
CQII.25	Which entities in the knowledge graph are linked to Wikidata, and what properties are mapped?	select	Aaron Warner
CQII.26	Which entities are classified as OOKB, meaning they have no matching entry in Wikidata or other authority sources?	select	Aaron Warner
CQII.27	Can I perform a federated SPARQL query combining Viewsari data with an external endpoint (e.g., Wikidata)?	descriptive	Aaron Warner

ID	Competency question	Type	Persona
CQII.28	How can I use Viewsari as a research tool for cross-disciplinary research that connects findings from *The Lives* to broader historical or cultural knowledge?	descriptive	Nazeera Marfi

Cluster 7 – User interaction and exploration

ID	Competency question	Type	Persona
CQII.29	How can I apply FAIR principles to the Viewsari data to bring it into my own project and share it with my students?	descriptive	Nazeera Marfi
CQII.30	How can I use Viewsari as an interactive visualization platform in my lectures to communicate complex relationships and historical patterns?	descriptive	Nazeera Marfi
CQII.31	Can I conduct a proof-of-concept project reusing Viewsari as a system, finding its working concept applicable to my own research?	descriptive	Nazeera Marfi
CQII.32	How can I integrate Viewsari into my teaching curriculum?	descriptive	Nazeera Marfi
CQII.33	How can I explore relationships between artists, artworks, or locations so that I can identify trends and accumulations of collaborations in a certain city or place?	descriptive	John Saffron

Cluster 8 – Extraction provenance and epistemic modeling

ID	Competency question	Type	Persona
CQII.34	Which extraction activity (NER run, LLM prompt run) generated entity E, and what software agent was used?	select	Aaron Warner
CQII.35	What was the prompt template used in the extraction run that produced entity E?	select	Aaron Warner
CQII.36	Which entities were extracted by a statistical method (co-occurrence / NER) versus a generative LLM-based method?	select	Aaron Warner
CQII.37	What model version and timestamp are associated with extraction run R?	select	Aaron Warner
CQII.38	Which entity mentions are typed as explicit and which as implicit, and can the two be retrieved separately?	select	Aaron Warner
CQII.39	For a given extraction result, can the provenance chain be traced back from the entity through the annotation, the paragraph, the page, the volume, and the translation to the original work?	select	Elena Rossi
CQII.40	Does the knowledge graph allow one to assess whether a claimed relationship between two artists is supported by explicit textual evidence, an implicit inference, or purely statistical co-occurrence?	select	Elena Rossi
CQII.41	How does the ontology represent the epistemic weight of a co-occurrence derived from PMI compared to an implicit mention inferred by an LLM?	descriptive	Aaron Warner

Summary: CQ coverage by cluster

#	Cluster	Phase I	Phase II	Total
1	Artists, collaborations, and relationships	CQI.1–8	CQII.1–5	13

#	Cluster	Phase I	Phase II	Total
2	Artworks	CQI.14–16	CQII.6–10	8
3	Locations	CQI.15–18	CQII.11–14	7
4	Historical events	–	CQII.15–18	4
5	Evidence and bibliographic information	CQI.19–23	CQII.19–24	11
6	Linked data enrichment	CQI.13, 27–28	CQII.25–28	7
7	User interaction and exploration	CQI.24–26, 29	CQII.29–33	9
8	Extraction provenance and epistemic modeling	–	CQII.34–41	8
	Persons and attributes (cross-cutting)	CQI.9–13	–	5
	Total	**29**	**41**	**70**

Cluster ranges overlap because some CQs address more than one cluster; the per-phase totals (29 Phase I, 41 Phase II, 70 overall) are the authoritative counts. The 74 raw Phase II questions generated via OntoChat (Section 7.3.2.1) are consolidated into the 41 Phase II catalog entries, with sub-questions subsumed under their parent entry.

SPARQL translations of the competency questions

This chapter mirrors parts of the CQ catalog. The complete set of executable SPARQL translations – one for every SPARQL-answerable competency question, used in the evaluation of Section 9.1.3 – is maintained in the project repository[1]: the machine-readable index lives in `src/evaluation/cq_catalog.json` and the executable `.rq` form of each query under `src/evaluation/queries/`. To keep the printed appendix compact, only the shared query

[1] The full SPARQL translations, the machine-readable index, and the evaluation scripts are versioned in the project repository at `https://github.com/ISE-FIZKarlsruhe/viewsari`, under the directory `src/evaluation/`.

infrastructure (the prefix block, the closure rules, and the provenance-coverage metric) and a small set of representative translations are reproduced below.

All queries share the prefix block in Listing A.1; per-query prefix declarations are omitted from the listings below for compactness. Queries that depend on the closure rules of Section 9.1.3 (the materialization of the per-instance `viewsari:0001032` edge on co-occurrence and person nodes) are flagged with a requires closure note.

Listing A.1: Shared SPARQL prefix block.

```
PREFIX viewsari: <https://viewsari.ise.fiz-karlsruhe.de/ontology/#>
PREFIX vkb:      <https://viewsari.ise.fiz-karlsruhe.de/kb/1.0#>
PREFIX prov:     <http://www.w3.org/ns/prov#>
PREFIX oa:       <http://www.w3.org/ns/oa#>
PREFIX doco:     <http://purl.org/spar/doco/>
PREFIX fabio:    <http://purl.org/spar/fabio/>
PREFIX frbr:     <http://purl.org/vocab/frbr/core#>
PREFIX dct:      <http://purl.org/dc/terms/>
PREFIX dc:       <http://purl.org/dc/elements/1.1/>
PREFIX foaf:     <http://xmlns.com/foaf/0.1/>
PREFIX skos:     <http://www.w3.org/2004/02/skos/core#>
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf:      <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl:      <http://www.w3.org/2002/07/owl#>
PREFIX xsd:      <http://www.w3.org/2001/XMLSchema#>
PREFIX wd:       <http://www.wikidata.org/entity/>
```

The three rules in Listing A.2 materialize the per-instance `viewsari:0001032` (paragraph anchor) edge that the closed evaluation graph depends on (cf. Section 9.1.3 and the discussion of the deployed–closed split in Section 8.3). They are applied in order so that R3 can chain off R2; the operation adds 7,431 derived triples to the 2,448,545-triple deployed dump (taking the closed graph to 2,455,976 triples; Section 8.3).

Listing A.2: Closure rules applied to the deployed Viewsari knowledge graph to produce the evaluation (closed) graph. R1 propagates the paragraph anchor down the artwork → annotation chain; R2 parses the cooccurrence `rdfs:label` to derive a direct cooccurrence → paragraph edge; R3 lifts that anchor onto the persons participating in each cooccurrence.

```
# R1: artwork closure via Web Annotation derivation
#   artwork -> wasDerivedFrom -> mention -> 0001032 -> paragraph
#   => artwork -> 0001032 -> paragraph
```

```
INSERT {
  ?entity viewsari:0001032 ?paragraph .
}
WHERE {
  ?entity prov:wasDerivedFrom ?mention .
  ?mention a oa:Annotation ;
           viewsari:0001032 ?paragraph .
  FILTER NOT EXISTS { ?entity viewsari:0001032 ?paragraph }
}  ;

# R2: cooccurrence closure (parse "Vol. N, Para. M" out of the label)
INSERT {
  ?cooc viewsari:0001032 ?paragraph .
}
WHERE {
  ?cooc a viewsari:0001025 ;
        rdfs:label ?label .
  FILTER (REGEX(?label, "Vol\\.\\s*\\d+,\\s*Para\\.\\s*\\d+"))
  BIND (REPLACE(?label, "^.*Vol\\.\\s*(\\d+),\\s*Para\\.\\s*(\\d+).*$", "$1") AS
    ↪ ?volNum)
  BIND (REPLACE(?label, "^.*Vol\\.\\s*(\\d+),\\s*Para\\.\\s*(\\d+).*$", "$2") AS
    ↪ ?paraNum)
  BIND (IRI(CONCAT(STR(vkb:the_lives_1568_volume-), ?volNum,
                   "_paragraph-", ?paraNum)) AS ?paragraph)
  FILTER NOT EXISTS { ?cooc viewsari:0001032 ?paragraph }
}  ;

# R3: person closure (chains off R2)
#   person <- involves <- cooccurrence -> 0001032 -> paragraph
#   => person -> 0001032 -> paragraph
INSERT {
  ?person viewsari:0001032 ?paragraph .
}
WHERE {
  ?cooc a viewsari:0001025 ;
        viewsari:0001034 ?person ;
        viewsari:0001032 ?paragraph .
  ?person a viewsari:0001013 .
  FILTER NOT EXISTS { ?person viewsari:0001032 ?paragraph }
}
```

The provenance-coverage metric (M06, Section 9.5.4) is computed by Listing A.3: it counts `oa:Annotation` nodes whose `prov:wasGeneratedBy` chain reaches a `prov:wasAssociatedWith` agent against the total, returning the closed-graph ratio (**112,113 / 112,113 = 1.0**). The two counts are computed in independent subqueries to avoid the `COUNT(DISTINCT)` miscounting that occurs when many annotations share one activity.

Listing A.3: Provenance coverage (metric M06): total `oa:Annotation` nodes versus those whose provenance chain reaches a documented `prov:wasAssociatedWith` agent, with the ratio.

```
PREFIX oa:   <http://www.w3.org/ns/oa#>
PREFIX prov: <http://www.w3.org/ns/prov#>

SELECT ?total ?withProv
       (IF(?total = 0, 0,
           ROUND(xsd:decimal(?withProv) * 10000.0 /
                 xsd:decimal(?total)) / 10000.0) AS ?ratio)
WHERE {
  {
    SELECT (COUNT(DISTINCT ?ann) AS ?total) WHERE {
      ?ann a oa:Annotation .
    }
  }
  {
    SELECT (COUNT(DISTINCT ?ann) AS ?withProv) WHERE {
      ?ann a oa:Annotation ;
           prov:wasGeneratedBy ?act .
      ?act prov:wasAssociatedWith ?agent .
    }
  }
}
```

Phase I translations

Artists and co-occurrences

```
SELECT DISTINCT ?a ?aLabel ?b ?bLabel WHERE {
  ?cooc a viewsari:0001025 ;
        viewsari:involves ?a, ?b .
  FILTER (str(?a) < str(?b))
```

```
  OPTIONAL { ?a rdfs:label ?aLabel }
  OPTIONAL { ?b rdfs:label ?bLabel }
}
LIMIT 200
```

```
SELECT DISTINCT ?paragraph ?a ?aLabel ?b ?bLabel WHERE {
  ?cooc a viewsari:0001025 ;
        viewsari:involves ?a, ?b ;
        viewsari:0001032 ?paragraph .
  FILTER (str(?a) < str(?b))
  ?a a viewsari:0001013 .
  ?b a viewsari:0001013 .
  OPTIONAL { ?a rdfs:label ?aLabel }
  OPTIONAL { ?b rdfs:label ?bLabel }
}
LIMIT 200
```

The remaining Phase I translations (`CQI.2`, `CQI.4`–`CQI.29`) follow the same query patterns illustrated above and are omitted here for brevity.

Phase II translations

Cluster 8 – Extraction provenance and epistemic modeling

The two translations below are reproduced as representative Phase II examples; they belong to Cluster 8, the cluster most directly tied to the dissertation's provenance and accountability claims. The remaining Phase II translations (Clusters 1–8, `CQII.1`–`CQII.38` and `CQII.41`) follow the same query patterns and are available as executable `.rq` files under `src/evaluation/queries/` in the project repository; the descriptive (non-SPARQL) CQs are answered as recorded in the catalog of the previous chapter (the *Type* column).

```
SELECT ?mention ?paragraph ?biography ?volume
       ?translation ?edition ?work WHERE {
  vkb:Q1069665 prov:wasDerivedFrom ?mention ;
               viewsari:0001032 ?paragraph .
  ?paragraph frbr:isPartOf ?biography .
  ?biography frbr:isPartOf ?volume .
  ?volume    frbr:isPartOf ?translation .
  OPTIONAL { ?translation viewsari:isBasedOn  ?edition }
  OPTIONAL { ?translation frbr:realizationOf  ?work    }
} LIMIT 5
```

```
SELECT ?evidenceType (COUNT(DISTINCT ?node) AS ?n) WHERE {
  {
    ?ma prov:wasDerivedFrom/^prov:wasGeneratedBy
        vkb:buonarroti-michelagnolo ;
        a viewsari:0001017 ;
        viewsari:0001032 ?p .
    ?mb prov:wasDerivedFrom/^prov:wasGeneratedBy
        vkb:sanzio-raffaello-raffaello-da-urbino ;
        a viewsari:0001017 ;
        viewsari:0001032 ?p .
    BIND (?p AS ?node) BIND ("explicit" AS ?evidenceType)
  } UNION {
    ?ma prov:wasDerivedFrom/^prov:wasGeneratedBy
        vkb:buonarroti-michelagnolo ;
        viewsari:0001032 ?p .
    ?mb prov:wasDerivedFrom/^prov:wasGeneratedBy
        vkb:sanzio-raffaello-raffaello-da-urbino ;
        viewsari:0001032 ?p .
    { ?ma a viewsari:0001016 } UNION
    { ?ma a viewsari:0001018 } UNION
    { ?mb a viewsari:0001016 } UNION
    { ?mb a viewsari:0001018 }
    BIND (?p AS ?node) BIND ("implicit/coreferent" AS ?evidenceType)
  } UNION {
    ?cooc a viewsari:0001025 ;
          viewsari:involves vkb:buonarroti-michelagnolo ,
                            vkb:sanzio-raffaello-raffaello-da-urbino .
```

```
    BIND (?cooc AS ?node)
    BIND ("statistical-cooccurrence" AS ?evidenceType)
  }
} GROUP BY ?evidenceType ORDER BY DESC(?n)
```

Each query reproduced above – and the full set in the repository – is the contents of an executable `.rq` file under `src/evaluation/queries/` (with the prefix block of Listing A.1 prepended). Running `python src/evaluation/run_cq_evaluation.py` executes all 51 SPARQL queries against `data/kg/viewsari_kg.inferred.ttl` and emits a coverage report; the results table that populates Section 9.1.3 is generated mechanically from the same JSON output.

Prompt catalog (ObliquER)

This appendix catalogs the prompt variants used in the ObliquER recognition and entity-linking pipeline. Each variant is documented along four dimensions, its purpose and design rationale, the content layers it activates, the structural template into which the content is rendered, and the corresponding source files in the project repository.[2]

The catalog distinguishes *recognition* prompts (P1–P3), used by ObliquER to identify and type artwork mentions in Vasari's text, from *entity-linking* prompts (P4–P5), used to match recognized mentions to Wikidata candidates. Of the latter two, P5 (few-shot ranker) is the production variant whose output populates the deployed knowledge graph; P4 (chain-of-thought) is retained for the empirical comparison in Chapter 9. Two ablation variants (A1–A2) are documented separately, corresponding to the conditions reported in Section 9.4.6.

[2] All prompt sources are versioned in the project repository at `https://github.com/ISE-FIZKarlsruhe/ObliquER/tree/main/data/viewsari/prompts`; the YAML content files (`prompt_parts.yaml`) and Jinja2 templates (`prompts.j2` and variants) referenced below are the canonical artifacts from which all rendered prompts are produced at inference time.

Output schema (recognition)

All recognition prompts produce JSON arrays in which every artwork mention is annotated according to the following base schema. The mention type vocabulary is shared across all variants.

Field	Description
`paragraph_id`	Identifier of the paragraph being annotated.
`mention_id`	Unique mention identifier of the form `m_XXXX`.
`type`	One of `explicit artwork mention`, `implicit artwork mention`, `coreferent`, or `generic collection mention`.
`surface_form`	Exact text span as it appears in the paragraph.
`entity_id`	Cluster identifier shared by all mentions of the same artwork.
`refers_to`	For coreferent mentions, the `mention_id` of the antecedent; `null` otherwise.
`in-text span annotation`	Surface form rendered in $\pm$5-token context, marked with « » brackets.

Of the prompt variants summarized in Table A.26, only the ontology-guided recognition variant (**P2**) is reproduced verbatim below as the representative case because we used it for the recognition output explicitly aligned to the Viewsari ontology classes (Section 7.3.2.2). The remaining variants and the two ablation conditions are documented and versioned in full in the ObliquER repository.[3]

Variant P2 – Ontology-guided

Purpose. Extends P1 by replacing the generic entity definition with one aligned to the Viewsari ontology classes (`viewsari:artwork` and the `viewsari:mention` subtypes from Section 7.3.2.2),

[3] All prompt sources are versioned at `https://github.com/ISE-FIZKarlsruhe/ObliquER/tree/main/data/viewsari/prompts`.

and adds a knowledge-graph modeling example showing the downstream PROV-O / Web Annotation representation that the recognition output feeds into. The rationale is twofold: grounding the model's output vocabulary in the ontology classes promotes label consistency, and the downstream representation provides additional signal for entity clustering across paragraphs.

Source files.

- `data/viewsari/prompts/few_shot/ontology_guided/prompts.j2`
- `data/viewsari/prompts/few_shot/ontology_guided/prompt_parts.yaml`

Content layers. all P1 layers, plus ontology-aligned entity definition referencing `viewsari:artwork`, `viewsari:explicit_artwork_mention`, `viewsari:implicit_artwork_mention`, `viewsari:coreferent`, and `viewsari:generic_mention`; KG modeling example showing the OA / PROV-O serialization of an extracted mention; two extended few-shot examples (Cimabue and Ghirlandaio).

Template (verbatim). The full Jinja2 template at `few_shot/ontology_guided/prompts.j2`. The `modeling_example` block injects the OA / PROV-O serialization that aligns the model's output with the downstream ontology.

```
{{ entity_definition }}

######
KNOWLEDGE GRAPH MODELING EXAMPLE
######

{{ modeling_example }}

{{ task }}

{{ special_cases }}

######
Provide the output as a JSON array in the following format:
######

[
{{ output_format }},
]

{{ input_description }}

{% for example in examples %}
######
EXAMPLE {{ loop.index }}{% if example.description %} -- {{ example.description }}{% endif %}

######
```

```
{{ example.paragraph }}
######
EXAMPLE OUTPUT {{ loop.index }}
######

{{ example.output }}

{% endfor %}
######
PREVIOUS PARAGRAPH
######

Paragraph {{ prev.paragraph_id }}):
{{ prev.text }}

######
OUTPUT
######

{{ prev.output }}

######
CURRENT PARAGRAPH
######

Paragraph {{ curr.paragraph_id }}):
{{ curr.text }}
```

Content layer overview

Table A.26 summarizes which content layers are activated in each variant.

All prompt variants are versioned in the ObliquER repository and rendered at inference time from the YAML content files and Jinja2 templates listed above. Running `python create_viewsari_prompts.py -variant <name>` produces the rendered `.j2` files for a given variant under `data/viewsari/prompts/<variant>/generated/`; each rendered file corresponds to one paragraph and is dispatched to the LLM API by `prompting.py`.

Table A.26: Content layers activated across the seven prompt variants. ✓ = layer present; – = layer absent.

Content layer	P1	P2	P3	P4	P5	A1	A2
Expert role description	✓	✓	✓	✓	✓	✓	–
Entity definition / taxonomy	✓	✓	✓	–	–	✓	–
Task instructions	✓	✓	✓	✓	✓	✓	✓
Special-case rules	✓	✓	✓	–	–	✓	–
Ontology alignment	–	✓	–	–	–	–	–
KG modeling example	–	✓	–	–	–	–	–
Explicit clustering rules	–	–	✓	–	–	–	–
Reasoning scaffold	–	–	✓	✓	–	–	–
Linking subtask in same pass	–	–	–	–	–	✓	–
Few-shot examples (count)	1	2	1	6	6	1	0

Knowledge graph serialization excerpt

The deployed Viewsari knowledge graph is distributed as a single Turtle file, `viewsari_kg.ttl` together with its inferred companion `viewsari_kg.inferred.ttl` produced by Population Run 1 and applying the closure rules of Listing A.2. Both files live under `data/kg/` in the project repository. Listing A.4 reproduces a small, hand-picked slice that traces one extracted artwork mention through the bibliographic, structural, content, and provenance layers, illustrating the chaining that the descriptive metrics M01–M07 (Table 8.8) quantify over the full graph. The paragraph body has been abbreviated to keep the excerpt readable; in the deployed file, the full text of every paragraph is carried verbatim in `viewsari:0001002`.

Listing A.4: Representative Turtle excerpt from `viewsari_kg.ttl`, showing one paragraph (structural), one artwork mention and its provenance chain (content + provenance), one co-occurrence with both participants (content), and the linked person individual with its Wikidata cross-reference (content). Paragraph text in `viewsari:0001002` is abbreviated with "...".

```
# -- Structural layer: paragraph anchor -----------------------------
vkb:the_lives_1568_volume-3_paragraph-265 a doco:Paragraph ;
    rdfs:label "The␣Lives,␣1568,␣Vol.␣3,␣Paragraph␣265" ;
    dct:isPartOf vkb:the_lives_1568_volume-3_ghirlandajo-bio_page_232 ;
    frbr:isPartOf vkb:the_lives_1568_volume-3_ghirlandajo-bio ;
    viewsari:0001005
        vkb:the_lives_1568_gutenberg_web_version_of_volume-3_ghirlandajo-
    ↪ bio_page_233 ;
    viewsari:0001004
```

```
        vkb:the_lives_1568_gutenberg_web_version_of_volume-3_ghirlandajo-
    ↪ bio_page_232 ;
    viewsari:0001003 "2323"^^xsd:nonNegativeInteger ;
    viewsari:0001002 "[...] they filled the whole arch of that chapel with angels
    ↪ , besides painting the folding doors that close the organ, and beginning
    ↪ to overlay the ceiling with gold. [...]" .

# -- Content layer: artwork mention + provenance chain --------------
vkb:108f2dd24ce043e898f5aae5de8181791060 a prov:Entity,
        viewsari:0001012, viewsari:0001033 ;
    rdfs:label "the whole arch of that chapel with angels" ;
    prov:wasDerivedFrom vkb:gt_ghirlandaio_m_0232 ;
    prov:wasGeneratedBy vkb:ground_truth_entity_linking_run_1 ;
    viewsari:0001032 vkb:the_lives_1568_volume-3_paragraph-265 .

vkb:gt_ghirlandaio_m_0232 a oa:Annotation, prov:Entity ;
    oa:hasBody vkb:body_gt_ghirlandaio_m_0232 ;
    oa:hasTarget vkb:textchunk_gt_ghirlandaio_m_0232 ;
    prov:wasGeneratedBy vkb:ground_truth_named_entity_recognition_run_1 ;
    viewsari:0001032 vkb:the_lives_1568_volume-3_paragraph-265 .

# -- Content layer: cooccurrence + participating persons ------------
vkb:cooccurrence_ghi-mai_1 a viewsari:0001025 ;
    rdfs:label "Co-occurrence of Ghirlandajo, Domenico and Mainardi, Bastiano,
    ↪ Vol. 3, Para. 265" ;
    prov:wasGeneratedBy vkb:cooccurrence_extraction_run_1 ;
    viewsari:0001034 vkb:ghirlandajo-domenico, vkb:mainardi-bastiano .

vkb:ghirlandajo-domenico a viewsari:0001013 ;
    rdfs:label "Ghirlandajo, Domenico" ;
    owl:sameAs wd:Q207681 .

# -- Provenance layer: typed extraction activity --------------------
vkb:ground_truth_named_entity_recognition_run_1 a prov:Activity,
        viewsari:0001040 ;       # viewsari:named_entity_recognition
    rdfs:label "Ground-truth NER run 1" ;
    prov:used vkb:the_lives_1568_volume-3_paragraph-265 .
```

List of Figures

1.1 Two complementary interpretive modes applied to the same historical text. Close reading (left) foregrounds the interpretive acts embedded in a single passage, and distant reading (right) reveals macro-structural patterns of co-mention across the full text that no single reader could perceive unaided. 5

1.2 Close reading of a text yields annotations, such as persons, dates, and locations, which are often limited to small-scale spaces. KE, however, infers content and adds semantic meaning. In comparison to the network in Figure 1.1, the graph here interrelates persons based on semantically meaningful connections, such as 'found in'. 6

1.3 René Magritte, *La trahison des images* (1929) 12

1.4 Naive KE applied to the Pietà authorship claim in Vasari's *The Lives*: the overheard attribution to Cristoforo Solari ("Our Gobbo from Milan") is encoded as a flat relation, with no record of its origin. 14

1.5 The same passage modeled in Viewsari's content layer: the explicit mention `Pietà` as a typed `oa:Annotation`, anchored to its source paragraph via `oa:hasTarget`, generated by a documented extraction run, and linked to `wd:Q235242` via `owl:sameAs`, where the actual creator (Michelangelo, `wd:Q5592`) is recorded. The contested Gobbo claim does not enter the KG as a creator triple. 14

2.1 Google's AI overview responding to the query *"how many rocks should i eat"* with a confident, sourced-sounding recommendation to consume at least one small rock per day for digestive health, attributed to UC Berkeley geologists. The response draws on a satirical article from *The Onion* that the model was unable to identify as such; a widely circulated example of questionable AI output presented with the epistemic authority of a verified fact (Sarkar 2025). 28

2.2 The same passage from Vasari's *Life of Leonardo* yields two distinct results depending on which interpretive convention the annotator foregrounds (Mailloux 1982); the provenance layer records the convention applied rather than resolving the ambiguity. 34

2.3 The ambiguity problem illustrated: a search for "Michelangelo" may retrieve results relating to the Teenage Mutant Ninja Turtle of that name rather than Michelangelo Buonarroti, the Renaissance sculptor and painter, and references to him in Giorgio Vasari's *The Lives*. Resolving such ambiguities requires EL grounded in contextual knowledge – which is only possible when textual sources are transformed into structured representations and the entities are modeled accordingly (see chapter 7). 40

3.1 The Semantic Web Technology Stack (after Nowack 2009), illustrating the layered architecture of W3C-standardized technologies that underpin the Semantic Web. From bottom to top: identification (URI/IRI), data exchange (XML, RDF), schema and ontology (RDFS, OWL, SKOS), rules (RIF), querying (SPARQL), and higher-level layers for logic, proof, trust, and security. 48

5.1 PROV-O representation of a named entity extraction step applied to a sentence from Vasari's *The Lives*. Solid purple boxes denote instances in the `viewsari` namespace; dashed green outlines denote the PROV-O classes to which those instances belong. 96

5.2 The `hico:InterpretationAct` is the single pivot of the representation: it is extracted from the source sentence (a `fabio:Expression`) via `hico:isExtractedFrom`, associated with the responsible human agent via `prov:wasAssociatedWith`, characterized by an interpretation type and criterion, and timestamped via `prov:atTime`. The claim produced by the act, here a `crm:E7_Activity` representing the papal commission event, is linked to the act via `prov:wasGeneratedBy`. Dashed outlines denote classes color-coded by namespace (green for PROV-O and FaBiO, amber for HiCO, blue for CIDOC CRM); solid purple boxes denote instances in the `viewsari:` namespace, and amber literal-style boxes hold the interpretation type and criterion values. 98

5.3 Example of how HiCO integrates with bibliographic and biographical vocabularies in a letter-corpus KG. A `pro:RoleInTime` binds a `foaf:Person` to a `pro:Role` over a specific `tvc:TimeInterval` and `bio:Place`, and is itself the product of a `hico:InterpretationAct` that records how a scholar derived the role assignment from a specific `fabio:Letter`. The interpretation act carries its own type and criterion (`hico:InterpretationType`, `hico:InterpretationCriterion`) and cites the source letter via `cito:citesAsSource` and `hico:isExtractedFrom`. Taken from Tomasi (2013). 100

5.4 The MEETUPs event model grounded in PROV-O. A MEETUP is modeled as a subclass of prov:Activity, produced by a Knowledge extraction activity (typed as an Algorithm) and derived from a Source (a prov:Entity) whose provider is classified as Wikipedia or another public web source. The extracted text content of the meetup is attached as an xsd:string literal via hasTextExtract. Adapted from Tirado et al. (2023). 101
5.5 A smell event modeled in the ODEUROPA ontology. A Smell Interaction (od:L15) is the result of a Smell Emission (od:L12) from one or more material sources (woodwork, leather, tapestry, camphor bags, potpourri), takes place in a gallery of an old house (here Knole), and motivates an Olfactory Experience (od:L13) in which the Author perceives a Smell (od:L11) characterized by multiple crm:E13_Attribute_Assignments (*dusty*, *sweeter*, *musty*, *old*). The model extends CIDOC CRM with the od: namespace for olfactory-specific classes and properties. Adapted from Lisena et al. (2022). 103
5.6 Harmonized view of the SEBI ontology instantiated on Lorenzo Valla's interpretation of the Donation of Constantine, combining SEBI's three subgraphs into a single representation. The evidence decomposes Valla's argument into its two orthogonal components, including a sebi:Feature being examined (philological arguments) and a sebi:Evaluation criterion under which the feature was judged (consistency). Adapted from Pasqual et al. (2024), Schimmenti et al. (2026). 105
5.7 Dual-provenance representation of a co-occurrence in Viewsari: paragraph 174 of Kloß's *Annalen der Loge zur Einigkeit* (1842) serves as the shared oa:hasTarget for a human FactGrid annotation and an automatic ObliquER extraction, each recorded as a distinct prov:Activity but converging on the same aligned entities (*Knigge*, *Lodge Zur Einigkeit*). The Viewsari instantiation of this schematic on real GT data is shown in Figure 10.6. 109
5.8 RDF reification of a weighted co-occurrence edge between Michelangelo and Raffaello in the social network, with the blank node _:e1 carrying the edge weight and :isEdgeOf links to both participants; per-node centrality measures (degree, betweenness, eigenvector) are attached as datatype properties, alongside foaf name assertions and an rdfs:seeAlso alignment to Wikidata. Adapted from Santini et al. (2024)'s website. 111

5.9 Excerpt from the WebVOWL visualization of the Viewsari ontology, generated via WIDOCO. Classes appear as blue circles and object properties as labeled arcs, with dashed lines indicating `rdfs:subClassOf` relationships to externally imported classes (here from FRBR and FaBiO). The excerpt shows the alignment between Viewsari's FRBR-derived core—`Work`, `Expression`, `Manifestation`, and `Item`—and the corresponding external classes. 115
5.10 Interactive D3.js rendering of the Viewsari ontology as embedded in the documentation portal. Classes are grouped and color-coded by modeling module (bibliographic, structural, content/entity, annotation, provenance, imported), with a side panel exposing the outgoing and incoming relations of the selected class—here `viewsari:cooccurrence`—and direct links to the corresponding WIDOCO entries. Complementing the WebVOWL view that ships with the HTML documentation, this custom visualization foregrounds the ontology's modular, provenance-aware structure. 116
5.11 Example of a SPARQL-driven result visualization in the Wikidata Query Service. The query (overlaid) retrieves all instances of paintings (`wdt:P31 wd:Q3305213`) whose creator (`wdt:P170`) is Gustav Klimt (`wd:Q34661`), along with their associated images (`wdt:P18`). The `#defaultView:ImageGrid` directive instructs the interface to render the results as a grid of thumbnails rather than a tabular listing, illustrating how lightweight view hints embedded in SPARQL queries can yield domain-appropriate presentations of CH data without bespoke interface code. 117
5.12 The *CoinSampo* portal, built with the Sampo-UI framework, presents a linked-data graph of Finnish numismatic finds. The left panel provides faceted filters for denomination, numismatic type, period, earliest/latest years, material, authority, mint, and findspot. The right panel offers interchangeable perspective tabs, such as table, map, heatmap, pie/bar chart, aoristic chart, migrations, minting-year animation, and export, all applied to the same filtered result set. The map displays the Finnish Heritage Agency's archaeological site polygons on an OpenStreetMap basemap, with individual coin finds shown as markers. Selecting a marker opens an entity card (Rantala et al. 2024b) 118

5.13 Entity landing page for the Belgian history painter Louis Gallait in the InTaVia visual analytics studio. The upper panel renders an ego-network of entities connected to Gallait (CH objects as green squares, events as green circles, places as purple triangles), while the lower panel arranges his attested events and productions along a horizontal timeline (c. 1835–1875). The interface exposes external authority links (Wikidata, GND, VIAF) and situates the actor within the InTaVia integrated KG of in/tangible European CH (Windhager et al. 2024). Screenshot taken from `https://intavia.acdh-dev.oeaw.ac.at/`. 120
6.1 Frontispiece of the second edition of Giorgio Vasari's *Le Vite de' Più Eccellenti Pittori, Scultori, e Architettori*, published in Florence by the Giunti press in 1568. The 1568 edition is the source text for the Viewsari corpus. Available via Wikimedia at `https://commons.wikimedia.org/wiki/File:Vite.jpg`. 122
6.2 Publication history of Vasari's biography of Leonardo da Vinci, showing the 1550 original Italian printing (left), the 1912 English translation (center), and the modern web-based digital version (right). 127
7.1 Distribution of PMI-based co-occurrences per volume in de Vere's translation of *The Lives*. 136
7.2 Distribution of Dice coefficient-based co-occurrences per volume in de Vere's translation of *The Lives*. 136
7.3 Top ten artists ranked by number of PMI-based co-occurrences across all volumes of de Vere's translation of *The Lives*. 137
7.4 Top ten artists ranked by number of Dice coefficient-based co-occurrences across all volumes of de Vere's translation of *The Lives*. 138
7.5 Overview of the three phases of the XD methodology, following the steps given in Blomqvist et al. (2016). The phases span project initiation and scoping, the iterative module development loop, and the integration loop. Outputs of each phase are shown below the respective phase. The iterative loop arrow indicates that newly released ontology versions feed back into the development process as new requirements emerge. 142
7.6 The Topic ODP, adapted from the ODP repository. A `Topic` mediates between a `Thing` (the resource being indexed) and a `Concept` (the semantic category serving as its subject), related through the inverse properties `isTopicOf` and `hasCore` (partially visible at the right edge of the figure). The pattern factors topical aboutness out of direct binary links, allowing additional metadata to be attached to the indexing relation itself. 148

7.7 Top-level classes of the Provenance ODP, adapted from the ODP repository and Shimizu et al. (2019): an `EntityWithProvenance` for which provenance is to be recorded, an `Agent` responsible for the assertion or generation, and a `ProvenanceActivity` representing the process through which the entity came about. The triad abstracts the core structure shared by PROV-O and related provenance vocabularies (Lebo et al. 2013). 148
7.8 The Participation ODP, adapted from the ODPs repository. An `Event` and an `Object` are related through the pair of inverse properties `hasParticipant` and `isParticipantIn`, with the class restriction `isParticipantIn some Event` expressing that any object qualifying as a participant must take part in at least one event. 149
7.9 Schematic overview of the Viewsari ontology. The three layers–Work, Instantiation, and Content–separate bibliographic identity (`:Book`/`:Edition`/`:Translation`) from physical-textual location (`:Volume`/`:Page`/`:Paragraph`) and from interpretive content (`:Biography`, `:Artist`, `:ArtistCooccurrence`). Artist co-occurrences are reified as first-class resources anchored to specific paragraphs, enabling provenance back to the textual locus of each relation. 150
7.10 Reification of a co-occurrence assertion via `rdf:Statement`. The triple (`:Raffaelo, :cooccursWith, :Michelangelo`) is reified as an `rdf:Statement` resource carrying its subject, predicate, and object as explicit properties, which permits attaching contextual metadata–here the textual locus (`:hasPage`, `:hasParagraph`) and the supporting evidence snippet (`textSnippet`)–without committing to the asserted triple itself. 153
7.11 Reification via the singleton property pattern. The instance-specific predicate `:cooccurrence` is declared an `rdf:singletonPropertyOf` the generic property `:cooccursWith` and used as the predicate of the triple (`:Raffaelo, :cooccurrence, :Michelangelo`), allowing contextual metadata such as `:hasParagraph` to be attached directly to the singleton predicate rather than to a separate reifying resource. 153
7.12 Reification via RDF-star. The asserted triple `:Raffaello :cooccursWith :Michelangelo` is itself referenced as a quoted triple (« . . . ») that serves as the subject of further statements attaching its textual locus (`:appearsIn`) and a PMI score (`:hasPMIScore`), avoiding the structural overhead of classical reification while preserving access to the original triple. 154

7.13 From untyped co-occurrence to typed relation. On the left, Michelagnolo is connected to other artists through undifferentiated co-occurrence edges that record textual adjacency without specifying its nature. On the right, a single pair is refined into a typed relation (`:collaboratedWith`). 155
7.14 Derivation of the co-occurrence pattern from the generic participation ODP. On the left, the participation pattern relates an `Event` and an `Object` through the inverse properties `hasParticipant` and `isParticipantIn`. On the right, the Viewsari specialization reifies the joint participation of two artists (Michelagnolo, Leonardo da Vinci) in a textual unit as a `:Co-occurrence` instance, linked to its participants via `:involvesResource`, anchored to its textual locus via `:appearsIn`, and quantified by a `hasPMIScore` value. 155
7.15 The OntoChat workflow (Zhang et al. 2024): an LLM-mediated pipeline from user-story elicitation through CQ extraction, analysis, and clustering, to ontology creation and testing, with domain experts and ontology engineers entering at distinct stages. 158
7.16 Persona-elicitation step of the OntoChat interface. A parameterized template guides the user in specifying a persona's name, age, occupation, skills, and interests, with explicit refinement criteria to ensure clarity and domain relevance. Buttons below grant access to templates for the remaining user-story components. The Markdown emphasis syntax visible in the slot placeholders (`**[NAME]**` and so on) reflects the fact that the domain experts authoring these templates are typically not experts in prompt engineering, surfacing a recurring tension in LLM-based ontology engineering between the conceptual literacy of the domain and the prompt-craft literacy presupposed by the tooling. 159
7.17 Worked example of how one paragraph of *The Lives* surfaces the requirement clusters that the Viewsari ontology has to satisfy, given the elicitation phase with domain experts. 160

7.18 The Viewsari bibliographic layer and its alignment with FaBiO and FRBR. Green dashed nodes represent classes reused from FaBiO and FRBR (`fabio:Work`, `fabio:Expression`, `fabio:Manifestation`, `fabio:Book`, `fabio:WebPage`, and related collection classes); pink dashed nodes represent Viewsari-defined subclasses that specialize these for Vasari's *Lives*, including `viewsari:translation` (a `fabio:Book`), the part-of hierarchy `viewsari:volume` → `viewsari:biography` → `viewsari:page`, and their HTML counterparts at the manifestation level. Solid black arrows denote FaBiO and FRBR object properties (`fabio:hasRealization`, `frbr:hasPart`, `fabio:hasEmbodiment`); dotted grey arrows denote `owl:subClassOf` relationships. 164
7.19 Bibliographic-layer example: Vasari's *Le Vite* modeled with FRBR/-FaBiO, from the abstract `fabio:Work` through the 1568 edition and de Vere translation down to volume, biography, and page-level embodiments, with the Project Gutenberg web manifestation as a parallel realization. 166
7.20 Structural-layer example: paragraph 103 of volume 1, with its plain-text content, character length, and embodiment as a `doco:TextChunk` on the Gutenberg web page. A named-entity annotation anchors the mention "Giotto" to the chunk via a `oa:TextPositionSelector`, with PROV-O recording the extraction activity, the extraction agent, and the index of names used as input. 167
7.21 Worked example of the extracted-content layer for the explicit artwork mention *S. Pietro Maggiore* in paragraph 284 of volume 3 of the de Vere translation. The mention is dual-typed as `oa:Annotation` and `prov:Entity`; the linked artwork entity is dual-typed as `viewsari:artwork` and `prov:Entity`. The two ObliquER pipeline runs that produced the mention and its Wikidata linking are reified as typed `prov:Activity` instances and connected to their software agents. 169
7.22 Worked example of the extracted-content layer for the artwork *Pietà* in paragraph 1412 of volume 9, with an additional `wd:P170` (*creator*) link from the artwork entity to `viewsari:michelangelo`. The dual-typing and provenance scaffolding are identical to Figure 7.21; only the domain enrichment differs. 170

7.23 Ontology-level representation of the mention-to-referent mapping in Viewsari. A textual mention (vkb:explicit_artwork_mention_1, typed as oa:Annotation, prov:Entity, and viewsari:explicit artwork mention) is linked via oa:hasBody to a body node carrying the surface form "Pietà". The body resolves to the artwork individual vkb:entity_0030 (typed as viewsari:artwork and prov:Entity), which is connected to its Wikidata counterpart via owl:sameAs. Linking provenance is captured by prov:wasGeneratedBy, pointing to the vkb:obliquer_entity-linking_run_1 activity (typed as prov:Activity), which in turn is associated with the software agent vkb:entity_linker_1 (typed as prov:SoftwareAgent) via prov:wasAssociatedWith. This separation makes recognition and linking provenance independently queryable at the ontology level. 185
7.24 Original GT sampling distribution ($N = 400$). Biographies are ordered chronologically by volume. Color encodes over-representation: red bars (> 30 paragraphs) indicate biographies whose share of the total corpus exceeds the per-biography cap adopted in the revised strategy. 194
7.25 Revised stratified GT distribution ($N = 208$, maximum 11% per biography, Michelangelo biography is an exception). All sixteen biographies are now fully annotated and retained in full (green bars). The only exception is Michelangelo, the anchor biography, for which 49 paragraphs were annotated during the schema-development phase; it is capped at 25 paragraphs ($\approx$12%) in the revised GT, with the gray extension indicating the 24 excess paragraphs excluded from the revised GT and reserved for the development and ablation sets. 195
7.26 INCEpTION annotation view for paragraphs 248–250 of the Ghirlandaio biography (de Vere Vol. III, pp. 219–221). Colored span labels indicate mention role: blue = Explicit Artwork Mention (with Wikidata QID where available), green/orange = Implicit Artwork Mention, yellow = Coreferent, purple = Generic Mention. Arcs labeled *Corefent Links* connect coreferent mentions to their antecedents. The paragraph header (# paragraph_id | pages | candidate_negative | text) is generated automatically from the merged-corpus CSV. 208

7.27 Distribution of annotated artwork mentions across the 16 biographies of the Viewsari GT corpus, ordered chronologically by Vasari volume (in parentheses below each label). Each bar is stacked by mention type – explicit artwork, implicit artwork, coreferent, and generic – and annotated with the per-biography total. The corpus-wide skew toward coreferent and implicit mentions is visible across biographies of all periods, with Michelangelo's biography ($N = 395$) and Brunelleschi's ($N = 272$) standing out both in absolute counts and in the relative weight of coreferent chains characteristic of Vasari's longer narratives. 212
7.28 The ObliquER pipeline. Stage 1 performs dynamic-chunked mention extraction with an LLM, typing spans as explicit, implicit, or coreferent. Stage 2 merges candidates from ReFinED, BLINK, GENRE, and an ArtPedia/SemArt-trained bi-encoder, then an LLM ranker selects a KB entity or assigns OOKB. 215
7.29 EL ranker as an LLM decision-maker. 235
7.30 Mapping the ArtPedia entry of Madonna (top) to the representation of an ObliquER paragraph-local mention record (bottom). The highlighted surface form is the sentence-level mention (either visual or contextual in the original), selected from the reconstructed paragraph; `type` carries the Viewsari mention type. 237
7.31 Explicit and implicit mention records for *The Scream*. 239
7.32 Example person node in the populated Viewsari KG, centered on `vkb:buonarroti-michelagnolo` (typed `viewsari:person`). Outgoing edges show the canonical label, the `owl:sameAs` link to `wd:Q5592`, three representative `skos:altLabel` variants harvested from Wikidata (with an ellipsis standing in for the remaining alternative spellings), and the `prov:wasGeneratedBy` link to the `vkb:coreference_reconciliation_run_1` activity. A second branch shows one mention annotation: the `oa:Annotation` resolves through `oa:hasTarget` to a `doco:TextChunk`, whose `oa:hasSelector` carries an `oa:TextPositionSelector` with explicit `oa:start`/`oa:end` character offsets, and whose `oa:hasSource` resolves back to the originating `viewsari:paragraph` instance. The annotation itself is provenance-linked to the `vkb:named_entity_recognition_run_1` activity. Literal values are drawn as rounded rectangles, IRI-typed individuals as ellipses, and external IRIs as dashed ellipses. 242

7.33 Provenance shape of an ObliquER per-paragraph NER sub-activity in the populated KG, instantiated on `vkb:ner_run_few_shot_v2_vol10_p024_p025`. The sub-activity (center) is typed both as `prov:Activity` and as `viewsari:named entity recognition`, and carries `prov:startedAtTime`/`prov:endedAtTime` from the inference log. `prov:wasInformedBy` links it upward to the strategy-level umbrella `vkb:ner_run_few_shot_v2`, which itself records the wall-clock span over the whole few-shot run and is associated, like the sub-activity, with the LLM agent `vkb:llm_agent_openai_gpt_oss_120b` (an `rdfs:seeAlso` link points to the model's Hugging Face page). Three `prov:used` edges record what the sub-activity consumed: the rendered prompt as a `prov:Entity` carrying the full prompt text on `rdfs:comment`, the context paragraph (here Vol. 10, paragraph 24), and the target paragraph (Vol. 10, paragraph 25). 245

8.1 Final form of the Viewsari bibliographic layer. The FaBiO/FRBR work *Le Vite* is connected to its 1568 Giunti edition, the 1912 de Vere English translation, and the Project Gutenberg manifestation collection, with dereferenceable web representations at the volume, biography, and page levels. 265

8.2 Viewsari content layer, with `viewsari:extracted_content` as the uniform PROV-typed superclass over persons, artworks, organizations, co-occurrences, and the mention taxonomy. Three typed activity classes (`named_entity_recognition`, `cooccurrence_analysis`, `entity_linking`) record the pipeline stage responsible for each instance. 267

8.3 Combined view of the content and bibliographic layers, joined via the structural layer. Every content-layer entity is anchored to a `doco:Paragraph`, which in turn carries page-extent properties pointing into the bibliographic layer's web representations, so that every extraction is traceable to a specific edition, translation, and online source page. 268

8.4 Landing page of the Viewsari web application, presenting corpus and KG statistics after Population Run 2, alongside the system's guiding design principle that knowledge extracted from interpretive texts is constructed rather than discovered, and must therefore model its own construction process. 284

8.5 The Viewsari knowledge base browser, showing aggregate statistics across all extraction runs (112,113 mentions, 15,804 artwork entities, 443 person entities, 632 co-occurrences) with a breakdown by source distinguishing GT annotations, ObliquER output, and other runs, alongside a filterable entity search interface. 285
9.1 Implicit artwork mention modeled in CIDOC CRM only. The panel is typed as `crm:E22_Human-Made_Object`, the production event as `crm:E12_Production`. The model asserts the artwork's existence as a real-world entity but provides no mechanism to record that it was *identified through an implicit reference* in the source text. The interpretive act is invisible. 294
9.2 Implicit artwork mention modeled in CIDOC CRM combined with PROV-O. Adding `prov:wasGeneratedBy` links the panel entity to an extraction activity, but this is semantically incoherent. The physical panel predates Vasari is not an output of any computational process. PROV-O's `wasGeneratedBy` conflates the *knowledge claim about* the artwork with the artwork itself. 294
9.3 Example instantiation of a Viewsari representation of an implicit artwork mention (the 'Last Supper'), showing two of the three representational layers, with a focus on the extraction process and result, as well as the process of linking the mention to a concrete artwork (the referent). 295
9.4 Co-occurrence extraction modeled in CIDOC CRMdig. The extraction run is typed as a `crmdig:D7_Digital_Machine_Event` whose input is a `crmdig:D1_Digital_Object` incorporating the Vasari 1568 edition (`crm:P165_incorporates`), and whose output is a `crmdig:D9_Data_Object`. 300
9.5 Reproduction of CRMinf specification example *A scholarly argument concerning Nero's whereabouts during the Great Fire of Rome* (Doerr et al. 2023), included under the CRMinf CC-BY 4.0 license. Francesca Bologna's argument is encoded as an `I1 Argumentation` decomposed into `I7 Belief Adoption` (from Tacitus), `I5 Inference Making` (rejecting the Suetonius–Cassius Dio account on plausibility grounds), and `I15 Provenance Assessment` (on the textual transmission of Suetonius). The entire chain is carried out by a single human `E39 Actor`. 303

9.6 A scholarly-fidelity CRMinf encoding of the Cimabue and Giotto co-occurrence claim (Life of Cimabue), inspired by and combining elements from the canonical CRMinf specification (Doerr et al. 2023). A fictional Vasari enthusiast adopts beliefs from two `E73 Information Objects` (Vasari's *The Lives* and a secondary critical commentary) via `I7 Belief Adoption` / `J7` / `J13`, performs an `S27 Observation` on Giotto's biography that expresses an `S28 Observable Situation` via `O36`, holds a prior `I2 Belief`, and synthesizes the four resulting premises through an `I5 Inference Making` (applying an `I3 Inference Logic`) into a new `I2 Belief` whose `I4 Proposition Set` is a CRM-formalized claim about a real-world event in late-thirteenth-century Florence. The entire argumentation framework is anchored in human-actor reasoning about real-world referents; there is no class for annotating text. Compare with the Viewsari encoding of the same passage in Figure 9.7. 306

9.7 Modeling the same passage in Viewsari. The interpretive act is cast as a `prov:SoftwareAgent` executing a typed `prov:Activity` that emits a `viewsari:Cooccurrence` annotation anchored via `oa:hasTarget` to the character-offset span in paragraph 12 of the de Vere translation of the 1568 Giuntina edition. The FRBR-aligned bibliographic chain (paragraph → biography → volume → translation → edition) preserves the exact editorial location of the textual mention. Compare with Figure 9.6. The LLM is neither a human actor nor a reasoner over referents, and the two cast choices are not interchangeable. 307

10.1 Worked example of the full three-layer Viewsari representation of a single Kloß *Annalen* extraction, taken from the mock KG. The coded alias *Philo* in § 174 is modeled across the bibliographic layer (the *Annalen* 1842 volume, `owl:sameAs` FactGrid Q11399, with its page and paragraph), the content layer (the `kx:coded_alias` annotation, its text chunk and position selector, and the resolved entity `owl:sameAs` FactGrid Q595), and the provenance layer (the typed extraction and identifier-recovery activities attributed to their respective agents). The structure is identical to the Vasari instantiation; only the domain content differs. 353

10.2 The resolved/OOKB bifurcation on two Kloß entity nodes (mock graph). *Top:* the coded alias *Philo* resolves to a person entity linked `owl:sameAs` FactGrid Q595 (Knigge), with its surface forms *Philo*, *Knigge*, and *Dieser* carried as `skos:altLabel`. *Bottom:* the explicitly named *Br. Herophilus* has no FactGrid or Wikidata referent and is retained as an OOKB entity with a stable internal IRI – the pipeline's discovery contribution for this corpus. Both nodes carry the same dual `oa:Annotation` / `prov:Entity` mention typing and full provenance. 354

10.3 The provenance of information represented in the ODEUROPA data model, illustrating how olfactory concepts (`od:L12 Smell Emission`, `od:L11 Smell`, `od:L13 Olfactory Experience`) relate to `crm:E73 Information Object` via `crm:P67 refers to`, and how provenance is captured in a separate layer linking the information object to `prov:Activity`, `prov:Agent`, `prov:SoftwareAgent`, and associated literals (e.g., confidence scores) through PROV-O properties. Figure as shown taken from Lisena et al. (2022) 357

10.4 The ANN pattern and its extension toward COMP-PROV, after Ondraszek et al. (2024c). Left: the abstract Annotation pattern – an annotation attached to a statement, recording provenance, purpose, place, time, and *agency*. Right: its Viewsari instantiation, in which the agency and provenance dimensions are realized with a PROV-O layer and a typed `prov:SoftwareAgent` – the merge that COMP-PROV formalizes for computationally-assisted extraction. Because the right-hand structure only adds a software agent to a PROV-O provenance chain, it applies directly to any DH KG already using PROV-O (e.g. ODEUROPA) and to the Kloß transfer (Figure 10.6). 357

10.5 The abstract COMP-PROV pattern. A `HumanAgent` (researcher/designer) `actsOnBehalfOf` a parametrized `ExtractionActivity` that a `SoftwareAgent` `associates` with; the activity `uses` an `Input` chunk and `generates` an `AnnotationOutput` – the mention layer – whose entity (`wasDerivedFrom` the output) resolves either to a KB referent or to an OOKB entity carrying a stable internal identifier. Holding this mention distinct from the referent is what keeps the resolution an attributable claim rather than a bare assertion. The attribute box records the run identifier, timestamp, prompt-template reference, and model configuration. 359

10.6 Mock-KG inspired rendering of the dual-provenance co-occurrence from the persons and organizations extracted with ObliquER (cf. the schematic in Figure 5.7). The co-occurrence of Knigge (`owl:sameAs` FactGrid Q595) and the Lodge Zur Einigkeit / L'Union (Q11161), asserted on § 174, is the shared `oa:hasBody` of two annotations with a common `oa:hasTarget`. A human FactGrid-editor annotation (left, attributed via `prov:wasAssociatedWith` to a `prov:Person`) and an ObliquER annotation (right, attributed to a `prov:SoftwareAgent` and recording the prompt as a used `prov:Entity`). The two layers compose without competing for the same slot. 360
10.7 Hypothetical instantiation of the COMP-PROV pattern in the ODEUROPA / *EOKG* context. The abstract roles bind to ODEUROPA-specific (fictional) classes, and the textual mention from which the concrete smell entity is later derived is marked as *AnnotationOutput*, anchored to the passage. `od:Smell` is then the resolved entity. 361
10.8 ODP for co-occurrence modeling. A `Co-occurrence` class (yellow circle) holds an `involves` object property ranging over `Object` instances (blue diamond), while each `Object` exposes an inverse `isParticipantIn` property linking back to an `Event`. The existential restriction *is involved in some Co-occurrence* (left, black arrow) captures the OWL 2 necessary condition that every participating object must be linked to at least one co-occurrence instance. 362

List of Tables

2.1 Example of a structured relational representation of artworks. 39
5.1 Comparison of human and GPT extractions on passages from *The Study of Song-Liao Diplomatic Relations*, taken from Wang and Zhang (2025). Failure modes include silent omission, overgeneralization to a single type, temporal misreading, and correct extraction when the surface form is explicit. 107
5.2 NER performance on passages from the *Geographical Records of the Liao History*. 107
6.1 Example of an Index of Names, an excerpt of the Index of the Craftsmen Mentioned in Volume II of the de Vere translation (Project Gutenberg, `https://www.gutenberg.org/files/25759/25759-h/25759-h.htm#Page_\protect\penalty\z@273`). 128
6.2 Structural statistics of the Viewsari corpus, computed from the extracted de Vere translation of *The Lives*. 129
6.3 Cross-page paragraphs per volume in the de Vere translation. A paragraph is counted as cross-page if its text body extends across at least two printed pages of the digital edition. 130
7.1 Coverage of entity types across the ExtrART reference dataset (corpus with 55 paragraphs, 27 chapters). 140
7.2 Requirements for Viewsari collected during the initial requirements elicitation phase. 146
7.3 External vocabularies reused by the Viewsari ontology, with the specific classes and properties used from each. Viewsari-defined terms inherit from these vocabularies via `rdfs:subClassOf` and `rdfs:subPropertyOf` axioms, so the ontology introduces no orphan vocabulary above the level of these imports. 177
7.4 New ontology elements introduced in the current version relative to the first formalization, with the motivating requirement and the pipeline stage that triggered each addition. 178

7.5 Biographies selected for the Viewsari GT corpus, ordered chronologically by artistic period and volume. *Rev.* gives the number of paragraphs in the GT corpus; *Neg* counts heuristic negative-example candidates; *Avg W* reports mean paragraph word count. Michelangelo Buonarroti (*Rev.* = 49) reflects the full prior annotation run rather than the 10–20 stratified cap applied to the remaining biographies (*). 197
7.6 Corpus-level statistics for the Viewsari GT as realized ($N = 232$ paragraphs, 16 biographies). The negative fraction is computed over the GT; the heuristic candidate-negative flag marks paragraphs likely to contain no artwork mention (see subsection 6.4.1). Volume distribution reflects Vasari's own three-part narrative of the progressive perfection of the arts. 199
7.7 Entity inventory for paragraph 58 of the Michelangelo biography (Volume 9, `michelangelo_enriched.json`, 11 mentions). All four mention types defined in Table 7.8 appear. Two coreferents (the statues, the tomb) point to antecedents introduced in earlier paragraphs and are resolved against the cross-paragraph cluster in the post-processing step of section 7.4.8.1. The generic mention of various sketches and designs is directly contrasted with the immediately following implicit mention, which singles out a single recoverable artwork from the generic set. Three surface forms (the designs and cartoons of the Last Judgment on the first wall, the work, it) anchor the coreference chain of `artwork_56`, demonstrating the implicit → coreferent transition that is the most frequent recognition pattern in the corpus. 202
7.8 Mention types in the ObliquER annotation schema. 204
7.9 Annotation fields for recognition records of artwork mentions. 205
7.10 Common coreference forms in Vasari's prose. 206

7.11 Per-biography breakdown of the Viewsari GT corpus by mention type. *Bios* stands for the artist biography; *Paras* for the number of paragraphs per biography. *Exp.*, *Imp.*, *Coref.*, and *Gen.* count, respectively, explicit artwork mentions, implicit artwork mentions, coreferent mentions, and generic mentions. *Mentions* is the row sum across all four types. *Entities* counts unique artwork `entity_id` clusters per biography; *LINKED* and *OOKB* report the knowledge-base coverage split between Wikidata-linked and OOKB entities. † The *Entities*, *LINKED*, and *OOKB* totals are raw per-biography counts taken before corpus-level deduplication (the 956 clusters reduce to 848 unique entities). Per biography, *LINKED* + *OOKB* can be smaller than *Entities*, because some clusters are left unresolved at annotation time, carrying neither a Wikidata link nor an OOKB flag; the 18 such clusters across the corpus account for the difference ($956 = 396 + 542 + 18$) and are assigned a resolution during deduplication. 213
7.12 Content layers activated across the ObliquER recognition prompt variants (Few-shot, Ontology-guided abbreviated as Onto., CoT) and the four ablation conditions A1–A4. ✓ = layer present; – = layer absent. A1 collapses recognition, linking, and coreference into a single prompt; A2 strips the domain-specific content layers down to a generic NER role and task; A3 keeps the ontology-typed entity definition of the ontology-guided variant but drops the KG modeling-example block and supplies no demonstrations; A4 varies only the demonstration count over the ontology-guided template. The empirical comparison of the seven configurations is reported in subsection 9.4.6. 217
7.13 Wikidata properties used for artwork entity representations in ObliquER. 233
7.14 Mapping from annotation record fields to Viewsari ontology constructs. The same schema is shared by the manually annotated GT and by the ObliquER recognition outputs of subsection 7.5.5. 243
8.1 Summary of Viewsari ontology elements by layer in the final version. New elements introduced in the current iteration relative to the first formalization are marked with †. 266
8.2 ObliquER prompt template inventory and KG deployment status. *In KG* indicates whether the variant's output is loaded into the deployed Viewsari KG (subsection 7.5.5); variants marked *No* are retained in the project repository for the empirical comparison in chapter 9 but do not contribute mentions or links to the populated graph. 269
8.3 Corpus-level statistics for the Viewsari ground truth ($N = 232$ paragraphs, 16 biographies). 270

8.4 Composition of the deployed Viewsari KG by source. The *Ground truth* column reports the manually annotated layer *as deployed* in the graph: 2,439 `prov:Person`-attributed mentions. This is the *pre-correction* ground truth from which the graph was populated; the annotated corpus was *subsequently* corrected down to 2,422 mentions by merging 17 headless-coreferent chains into their antecedents (section 7.4.7). Therefore, the GT layer deployed and reported in this table is different from the corrected corpus in chapter 7 and will be reconciled in the next redeployment. The *ObliquER extraction* column is the recognition layer over all biography instances and paragraphs, and *KG (both)* is their union; all three columns report the canonical build (population Run 2). Mention rows are additive (ground truth + ObliquER = both); the biography and paragraph rows are not, because the ground truth is the 16-biography / 232-paragraph annotated subset of the full corpus. The artwork-entity rows are likewise additive: Run 2 links and clusters the recognized artworks of the re-authored ontology-guided variant (subsection 8.3.5), so the ObliquER entity column is now populated rather than pending. The GT column reports the deployed pre-correction layer (852 globally-deduplicated artworks; the corrected corpus of section 7.4.7 resolves to 848, and the per-biography-sum convention used in the accompanying conference paper yields 956). The remaining two recognition variants – few-shot and the original ontology-guided – contribute mentions but are not yet clustered, so the ObliquER entity counts are an upper bound that a global reconciliation pass will reduce (subsection 8.3.2). subsection 6.4.2 reports 168 biographies as the source-corpus scope; of these, 165 are materialized under `viewsari:biography`, the 3-unit gap reflecting supplementary entries in the de Vere translation (the academicians-of-design and festive-preparations appendices and Vasari's self-life in volume 10). 271

8.5 Raw mention counts per prompt variant from the Run 1 ObliquER recognition stage over all 10 volumes of the de Vere translation (`openai/gpt-oss-120b`). Counts are aggregated from the per-paragraph response files. These are the three prompt-engineering variants from which Run 1's ingested layer was drawn: the few-shot and ontology-guided variants as parallel mention layers, while the chain-of-thought variant is retained in the repository but loaded into neither build. Run 2 additionally ingests a re-authored ontology-guided variant (28,969 mentions; per-type composition in the ObliquER column of Table 8.4) not shown here, taking the canonical ObliquER layer to 84,215. 272

8.6 Mention and entity statistics for the ArtPedia ground truth as used in this dissertation (Wikipedia-paragraph reconstruction of the dataset). Annotations are restricted to *explicit* and *implicit* mention types; coreferent and generic annotations are not used in this split. OOKB counts are reported as percentages of the total number of entities and mentions. 272
8.7 Raw mention counts from the ObliquER recognition stage over ArtPedia (2,679 paintings; `openai/gpt-oss-120b`). Counts are aggregated per-painting; parsing failures are excluded (9 of 2,679 in the few-shot variant; 92 in the zero-shot variant). 273
8.8 Descriptive population metrics: what is in the populated Viewsari KG. Each metric is computed by a single SPARQL query; the dissertation cross-reference column points to the subsection where the result is reported. 277
8.9 Instance counts by Viewsari ontology class after population Run 2 (first stage + ObliquER second stage, with the new ontology-guided variant clustered into the entity layer) on the closed KG (2,455,976 triples). The Run 1 column carries the prior build for comparison (subsection 8.3.5); rows in the bibliographic and structural layer are identical across the two builds. 278
8.10 Mention taxonomy breakdown and instantiation count per mention type. 280
8.11 Activity inventory (M04) for population Run 2, with the Run 1 build for comparison. Each `prov:Activity` is typed by its Viewsari activity subclass. Whereas Run 1 recorded the linking pass as a single long-running corpus-level activity, Run 2's ontology-guided linking variant records entity linking *per paragraph*, raising the `entity_linking` count from one to 3,100; the two typed classes sum to 13,160, 634 short of the 13,794 `prov:Activity` total of Table 8.9, the difference being the 632 per-co-occurrence extraction activities together with the legacy foundation activities that carry no recognition/linking subclass. 280
8.12 Population Run 2 (the canonical build) measured against the prior Run 1. Both columns report measured counts from the respective built graph; the Run 2 column is the source of every KG figure reported in this dissertation, and the Run 1 column is retained only to document the change. Triple counts are over the prompt-stripped serialization (section 8.3); the GT layer is the pre-correction layer in both builds. 283
8.13 Functional requirements coverage by Viewsari interface view. 286

9.1 CQ coverage of the Viewsari ontology (closed KG with 2,455,976 triples). Percentages in the SPARQL-only column are calculated across the 52 SPARQL-translatable CQs; percentages of the total are calculated across all 70 catalog entries. 290
9.2 SPARQL query complexity distribution. 291
9.3 CQ coverage per thematic cluster (post-closure). 292
9.4 Comparative representational capabilities. CIDOC CRM, the CIDOC CRM + PROV-O + Web Annotation stack, and Viewsari across six dimensions relevant to generative, provenance-aware extraction from interpretive humanities texts. 299
9.5 Bootstrap 95 % confidence intervals on the strictest-criterion F1 for every recognition configuration reported on the Viewsari GT in this chapter. The *Strict* criterion (exact span and type) under surface-form matching and the *Strict* criterion (exact span and type) under span-level bipartite matching, computed via per-paragraph resampling with $N = 10{,}000$ replicates following Dror et al. (2018). 313
9.6 Strict and partial surface-form matching for ObliquER versus zero-shot NER baselines on the Viewsari GT and on ArtPedia. *Strict* = exact span and type; *Partial* = half credit for overlapping spans regardless of type. Viewsari rows are broken down by mention subset (Explicit, Implicit, All); ArtPedia is reported overall, since its annotation schema collapses to two mention types. The Viewsari ObliquER rows report the production ontology-guided 2-shot configuration under `openai/gpt-oss-120b`; the UniNER-7B and GLiNER2 baseline rows are re-run on the same unified-exclusion partition as the ObliquER rows. See the partition footnote on the lead paragraph of this subsection for the evaluation-set definition shared across all Viewsari cells of this chapter. 317
9.7 ObliquER entity-recognition results under *span-level bipartite matching*, decomposed across the exact, partial, and type criteria. *Strict* requires an exact span and exact type; *Partial* gives character-overlap credit regardless of type; *Type* gives character-overlap credit with matching type. The Viewsari columns report the production ontology-guided 2-shot configuration under `openai/gpt-oss-120b`; the ArtPedia columns report the few-shot configuration. 318
9.8 Per-mention-type partial recognition behavior of ObliquER on the Viewsari GT. Rows: gold types; columns: prediction. Generic and Coreferent gold rows are excluded because they were not evaluated (out of scope). Computed on the updated Viewsari GT under the unified-exclusion partition (see the footnote of subsection 9.4.3). 318

9.9 Per-mention-type partial recognition behavior of the few-shot configuration on the ArtPedia evaluation set. Rows: gold; columns: prediction. The ArtPedia annotation schema collapses gold-side to two mention types (Implicit, Explicit), with no Generic or Coreferent gold rows; the prediction side carries the full Viewsari four-way typology plus a *Missed* column for gold mentions for which the model produced no overlapping span. 319
9.10 Per-biography recognition behavior of the production ontology-guided 2-shot configuration on the Viewsari GT under `openai/gpt-oss-120b` and span-level bipartite matching, ranked by Type F1 (descending). The n column gives the gold-mention count per biography. The bottom row reproduces the chapter-wide aggregate. 320
9.11 EL on the Viewsari GT under gold-mention inputs, broken down across all, explicit, and implicit mention subsets. P/R/F1 are reported separately for KB-linked and OOKB mentions; CR denotes candidate recall. GENRE is generative end-to-end and does not use a candidate set, so CR is reported as "–". 321
9.12 EL on the ArtPedia test split under gold-mention inputs, broken down across All, Explicit, and Implicit mention subsets. P/R/F1 are reported separately for KB-linked and OOKB mentions; CR denotes candidate recall. GENRE is generative end-to-end and does not use a candidate set, so its CR row reports the post-decoding correctness rate rather than a retrieval-set candidate recall. 323
9.13 End-to-end EL on the Viewsari GT: mentions produced by the ObliquER NER stage feed directly into the EL stage, so the reported numbers reflect both recognition and linking errors. P/R/F1 are reported separately for KB-linked and OOKB mentions; CR denotes candidate recall. 324
9.14 Input-representation ablation. Candidate recall (CR@16) for the two dense retrievers under two query encodings, evaluated on the ArtPedia test partition. *Structured* = the LLM-generated structured metadata record used as the query; *Context* = the raw mention embedded in its surrounding paragraph context. 325

9.15 Linking floor (A1 monolith) vs. the production ObliquER linking pipeline on the Viewsari GT. A1 collapses recognition, linking, and coreference into one prompt and produces Wikidata identifiers from the LLM's parametric memory alone, without retrieval. CR is not applicable to A1 (no candidate set). The two ObliquER rows give the gold-mention and end-to-end settings, respectively, reproduced from Table 9.11 (*All* row) and Table 9.13 (*All* row); A1 is necessarily end-to-end because it emits its own mentions. The *n* column records the mention basis of each row. A1 evaluates on its own 1,422 emitted mentions (416 explicit + 1,006 implicit), the ObliquER end-to-end row on its own NER-emitted mentions, and the ObliquER gold-mention row on the 1,010 explicit-and-implicit mention instances of the GT. All A1 values extend the published evaluation. 328

9.16 Cross-model recognition ablation on Viewsari, ontology-guided template held constant, evaluated on the updated Viewsari GT (subsection 7.4.4). Recall (R) and F1 are reported under span-level bipartite matching across the Type, Partial, and Strict criteria for two inference models – `openai/gpt-oss-120b` and `google/gemma-4-31b-it` – at five demonstration counts $k \in \{0, 1, 2, 3, 5\}$. † marks the primary configuration used in the main results of Table 9.6. Both models served from the GWDG *chat-AI* endpoint with identical sampling parameters; *Precision* is omitted to keep the table compact (full P/R/F1 for the $k = 2$ rows is in Table 9.7 and Table 9.17). 330

9.17 Architectural ablation (A1–A3) on the Viewsari GT under span-level bipartite matching, reported under both inference models `openai/gpt-oss-120b` and `google/gemma-4-31b-it` (Strict P/R/F1 plus headline F1 under Partial and Type). The ObliquER reference rows reproduce the corresponding $k = 2$ entries of Table 9.16. 335

9.18 Few-shot count ablation (A4) on the Viewsari GT under span-level bipartite matching, ontology-guided template held constant for `gpt-oss-120b` (subsection 7.4.4). *Strict* and *Partial* are the span-level criteria of Table 9.7. The $k = 0$ row matches the GPT A3 row of Table 9.17; the $k = 2$ row is the GPT production reference. The demonstration-paragraph column is cumulative. Each entry lists the paragraphs added relative to the row above. The corresponding `gemma-4-31b-it` curve is reported in Table 9.16. All rows are evaluated on the unified-exclusion test partition described in the preamble of this subsection. 337

9.19 Prompt-variant comparison on the Viewsari GT under `openai/gpt-oss-120b` and span-level bipartite matching, evaluated on the unified-exclusion test partition. Each row varies the prompt structure while holding the chunking strategy, the evaluation suite, and the sampling parameters constant. The Ontology-guided ($k = 2$) row reproduces the GPT production reference of Table 9.17. 337
9.20 Catalog of evaluative KG metrics, the section in which each is reported, and its pass criterion. Descriptive counterparts (M01–M04, M08, M10, M11) are reported in section 8.3. 340
9.21 External linking coverage by entity type on the populated KG (metric M09). 341
9.22 Web-manifestation reachability per expression-level node (metric M12). 342
10.1 Summary statistics for the Kloß *Annalen* annotated corpus. 350
10.2 ObliquER recognition on the 64-paragraph Kloß *Annalen* GT (236 scored gold mentions), under the single adapted few-shot prompt held constant across two inference models. F1 is reported under the SemEval 2013-9.1 span-level criteria (*Strict* = exact span, exact type; *Partial* = half credit for overlapping spans; *Type* = correct type with overlapping span). 352
10.3 Updated comparative pattern analysis across four DH case studies, extended with the COMP-PROV pattern introduced in this dissertation. ✓ = pattern explicitly employed; (✓) = pattern implicitly instantiated or applicable; × = pattern absent. Abbreviations: Situation (*SIT*), Description (*DESC*), Visual Semiotics (*VIS-SEM*), Interpretation (*INT*), Information Realization (*IR*), Provenance (PROV), Person (*PER*), Place (*PL*), Annotation (ANN), Analog vs. Digital (*AN-DIG*), Computational Provenance (COMP-PROV). 358
12.1 Mapping of research questions to contributions and evidence 392
A.26 Content layers activated across the seven prompt variants. ✓ = layer present; – = layer absent. 424

List of Publications

All URLs and DOIs cited in this work were last accessed on June 14, 2026.

Felix Bach, Sandra Göller, Timo Holste, and Sarah Rebecca Ondraszek. Forschungsdaten suchen, finden und nachnutzen: Der NFDI4Memory Data Space als föderierte Infrastruktur für die historisch arbeitenden Geisteswissenschaften. In *FORGE 2025 - Forschungsdaten in den Geisteswissenschaften: Daten neu denken. Konferenzabstracts*. Zenodo, September 2025. doi: 10.5281/zenodo.17178216.

Oleksandra Bruns, Tabea Tietz, Linnaea Söhn, Jonatan Jalle Steller, Sarah Rebecca Ondraszek, Etienne Posthumus, Torsten Schrade, and Harald Sack. What's Cooking in the NFDI4Culture Kitchen? A KG-based Research Data Integration Workflow. In *Proceedings of the 4th Workshop on Metadata and Research (objects) Management for Linked Open Science (DaMaLOS 2024) at the Extended Semantic Web Conference (ESWC 2024)*, Hersonissos, Crete, Greece, May 2024. PUBLISSO Fachrepositorium. doi: 10.4126/FRL01-006474028. URL `https://repository.publisso.de/resource/frl:6474028`.

Sandra Göller, Timo Holste, Sarah Rebecca Ondraszek, and Tabea Tietz. The NFDI4Memory Data Space and RADAR4Memory. In *Proceedings of the Conference on Research Data Infrastructure (CoRDI 2025)*, Aachen, DEU, August 2025. doi: 10.5281/zenodo.16736275.

Sarah Rebecca Ondraszek. Viewsari: Enabling New Perspectives on the Renaissance with a Knowledge Graph of Giorgio Vasari's *The* Lives. In *Proceedings of the ISWC 2025 Companion Volume*, volume 4085 of *CEUR-WS*, Nara, Japan, November 2025. CEUR-WS.org. URL `https://ceur-ws.org/Vol-4085/paper17.pdf`.

Sarah Rebecca Ondraszek, Grischka Petri, Ulrike Blumenthal, Lisa Dieckmann, Etienne Posthumus, and Harald Sack. eXtreme Design for Ontological Engineering in the Digital Humanities with Viewsari, a Knowledge Graph of Giorgio Vasari's *The* Lives. In *Proceedings of the 1st International Workshop of Semantic Digital Humanities (SemDH 2024)*, volume 3724 of *CEUR-WS*, Hersonissos, Greece, May 2024a. CEUR-WS.org. URL `https://ceur-ws.org/Vol-3724/paper5.pdf`.

Sarah Rebecca Ondraszek, Harald Sack, and Etienne Posthumus. Viewsari: New Perspectives on Historical Network Analysis in Giorgio Vasari's *The* Lives Using Knowledge Graphs. In Martin Grandjean, editor, *Historical Network Research Conference 2024 — Book of Abstracts*, pages 137–139, Lausanne, 2024b. Zenodo. doi: 10.5281/zenodo.12665443.

Sarah Rebecca Ondraszek, Bruno Sartini, Marieke van Erp, Andrea Poltronieri, Pasquale Lisena, and Harald Sack. One pattern to express them all? Towards Generalized Patterns for Ontology Design in the Digital Humanities. In *Proceedings of Workshop on Ontology Design and Patterns (WOP 2024)*, Baltimore, MD, United States of America, November 2024c. Zenodo. doi: 10.5281/zenodo.14063917. URL https://www.fiz-karlsruhe.de/sites/default/files/FIZ/Dokumente/Forschung/ISE/Publications/Conferences-Workshops/WOP%40ISWC2024_OnePattern_ondraszek.pdf.

Sarah Rebecca Ondraszek, Tabea Tietz, Etienne Posthumus, Heike Fliegl, Jörg Waitelonis, and Harald Sack. Indexing Historical Research Data: MemO and the NFDI4Memory Knowledge Graph. In *Proceedings of the Conference on Research Data Infrastructure (CoRDI 2025)*, Aachen, DEU, August 2025a. doi: 10.5281/zenodo.16736125.

Sarah Rebecca Ondraszek, Tabea Tietz, Jörg Waitelonis, and Harald Sack. Interlinking Research Data and Services in the Historical Sciences with MemO and the NFDI4Memory Knowledge Graph. In *Proceedings of the 5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment*, Nara, Japan, October 2025b. CEUR Workshop Proceedings (CEUR-WS.org). doi: 10.5281/zenodo.17376359. URL https://ceur-ws.org/Vol-4065/paper09.pdf.

Sarah Rebecca Ondraszek, Jörg Waitelonis, Katja Keller, Claudia Niessner, Anna M. Jacyszyn, and Harald Sack. Ontologies in Motion: A BFO-Based Approach to Knowledge Graph Construction for Motor Performance Research Data in Sports Science. In *Proceedings of the 5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment*, volume Vol-4065 of *CEUR-WS*, pages 97–106, Nara, Japan, 2025c. CEUR Workshop Proceedings. URL https://ceur-ws.org/Vol-4065/paper08.pdf.

Bibliography

All URLs and DOIs cited in this work were last accessed on June 14, 2026.

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. doi: 10.1109/ACCESS.2018.2870052.

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. Entity linking via explicit mention-mention coreference modeling. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4644–4658, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.343. URL `https://aclanthology.org/2022.naacl-main.343/`.

Manfred Agethen. *Geheimbund und Utopie. Illuminaten, Freimaurer und deutsche Spätaufklärung*. Number 11 in Ancien Régime, Aufklärung und Revolution. R. Oldenbourg, München, 1984.

Annastiina Ahola, Lilli Peura, Rafael Leal, Heikki Rantala, and Eero Hyvönen. Using generative AI and LLMs to enrich art collection metadata for searching, browsing, and studying art history in digital humanities. In *Proceedings of the 2nd International Conference on Data & Digital Humanities: Generative Artificial Intelligence for Text and Multimodal Data*, University of Minho, Braga, Portugal, December 2024. URL `https://seco.cs.aalto.fi/publications/2024/ahola-et-al-genai-2024.pdf`.

Annastiina Ahola, Heikki Rantala, and Eero Hyvönen. ArtSampo – finnish art on the semantic web. In Albert Meroño Peñuela, Oscar Corcho, Paul Groth, Elena Simperl, Valentina Tamma, Andrea Giovanni Nuzzolese, Maria Poveda-Villalón, Marta Sabou, Valentina Presutti, Irene Celino, Artem Revenko, Joe Raad, Bruno Sartini, and Pasquale Lisena, editors, *The Semantic Web: ESWC 2024 Satellite Events*, page 159–163, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-78952-6.

Laura Albers, Peggy Große, and Sarah Wagner. *Semantic Data-Modeling and Long-Term Interpretability of Cultural Heritage Data—Three Case Studies*, page 239–253. Springer International Publishing, Cham, 2020. ISBN 978-3-030-15200-0. doi: 10.1007/978-3-030-15200-0_16. URL https://doi.org/10.1007/978-3-030-15200-0_16.

Dean Allemang and James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann, Waltham, MA, 2nd edition, 2011. ISBN 978-0-12-385965-5.

Elvira Amador-Domínguez, Emilio Serrano, Daniel Manrique, Patrick Hohenecker, and Thomas Lukasiewicz. An ontology-based deep learning approach for triple classification with out-of-knowledge-base entities. *Information Sciences*, 564:85–102, July 2021. ISSN 0020-0255. doi: 10.1016/j.ins.2021.02.018. URL https://www.sciencedirect.com/science/article/pii/S0020025521001602.

Diego Raphael Amancio. Network analysis of named entity co-occurrences in written texts. *EPL (Europhysics Letters)*, 114(5):58005, 2016. ISSN 1286-4854. doi: 10.1209/0295-5075/114/58005. URL http://dx.doi.org/10.1209/0295-5075/114/58005.

Anthropic. The Claude 3 model family: Opus, sonnet, haiku. Model Card, 2024. URL https://www.anthropic.com/claude.

Anthropic. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet, February 2025.

Dörthe Arndt, Pierre-Antoine Champin, et al. RDF-star and SPARQL-star. W3c working group note, W3C, 2022. URL https://www.w3.org/2022/08/rdf-star-wg-charter.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Mag.*, 36(1):15–24, March 2015. ISSN 0738-4602. doi: 10.1609/aimag.v36i1.2564. URL https://doi.org/10.1609/aimag.v36i1.2564.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The*

Semantic Web, page 722–735, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-76298-0.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. Generalisation in named entity recognition: A quantitative analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 524–533. Association for Computational Linguistics, 2017.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In Anastassia Loukina, Rashmi Gangadharaiah, and Bonan Min, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-industry.24. URL https://aclanthology.org/2022.naacl-industry.24/.

Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter Patel-Schneider. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge, 2003. ISBN 978-0-521-78176-6.

Leonardo Banh and Gero Strobel. Generative artificial intelligence. *Electronic Markets*, 33(1): 63, December 2023. ISSN 1422-8890. doi: 10.1007/s12525-023-00680-1. URL https://doi.org/10.1007/s12525-023-00680-1.

Barcelona Supercomputing Center. BSC and Prado Museum teach AI to view and interpret works of art, June 2023. URL https://www.bsc.es/news/bsc-news/bsc-and-prado-museum-teach-ai-view-and-interpret-works-art.

Michael Baxandall. *Giotto and the Orators: Humanist Observers of Painting in Italy and the Discovery of Pictorial Composition, 1350–1450*. Clarendon Press, Oxford, 1971.

Michael Baxandall. *Patterns of Intention: On the Historical Explanation of Pictures*. Yale University Press, New Haven, 1985.

Christin Beck and Marisa Köllner. GHisBERT – training BERT from scratch for lexical semantic investigations across historical German language stages. In Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen, David Alfter, Francesco Periti, and Pierluigi Cassotti, editors, *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 33–45, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.lchange-1.4. URL https://aclanthology.org/2023.lchange-1.4/.

David Beckett, Tim Berners-Lee, Eric Prud'hommeaux, and Gavin Carothers. RDF 1.1 turtle: Terse RDF triple language. W3c recommendation, W3C, 2014. URL https://www.w3.org/TR/turtle/.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 610–623, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

Kamal Berahmand, Yuefeng Li, and Yue Xu. A deep semi-supervised community detection based on point-wise mutual information. *IEEE Transactions on Computational Social Systems*, 11 (3):3444–3456, 2024. doi: 10.1109/TCSS.2023.3327810.

Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284 (5):34–43, 2001. doi: 10.1038/scientificamerican0501-34.

Ernst Bernheim. *Lehrbuch der historischen Methode und der Geschichtsphilosophie. Mit Nachweis der wichtigsten Quellen und Hilfsmittel zum Studium der Geschichte*. Duncker & Humblot, Leipzig, 5. und 6., neu durchgearbeitete und vermehrte aufl. edition, 1908.

Bibliothèque nationale de France. Gallica: Bibliothèque numérique, 1997. URL https://gallica.bnf.fr.

Ahsan Bilal, David Ebert, and Beiyu Lin. LLMs for Explainable AI: A Comprehensive Survey, March 2025. URL http://arxiv.org/abs/2504.00125. doi: 10.48550/arXiv.2504.00125.

Kevin Blissett and Heng Ji. Cross-lingual NIL entity clustering for low-resource languages. In Maciej Ogrodniczuk, Sameer Pradhan, Yulia Grishina, and Vincent Ng, editors, *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 20–25, Minneapolis, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2804. URL https://aclanthology.org/W19-2804/.

Eva Blomqvist, Karl Hammar, and Valentina Presutti. Engineering Ontologies with Patterns – The eXtreme Design Methodology. In *Ontology Engineering with Ontology Design Patterns*, pages 23–50. IOS Press, 2016. doi: 10.3233/978-1-61499-676-7-23. URL https://ebooks.iospress.nl/doi/10.3233/978-1-61499-676-7-23.

Gerd Blum. *Giorgio Vasari: der Erfinder der Renaissance*. C.H. Beck, München, 2011. ISBN 978-3-406-69954-2.

Anthony Blunt. *Artistic Theory in Italy, 1450–1600*. Clarendon Press, Oxford, 1962. Paperback reissue; first edition 1940.

BNC Consortium. The british national corpus, XML edition, 2007. URL http://www.natcorp.ox.ac.uk/.

Ksenija Bogetić, Vuk Batanović, and Nikola Ljubešić. Corpus compilation for digital humanities in lower-resourced languages: A practical look at compiling thematic digital media corpora in serbian, croatian and slovenian. *Suvremena lingvistika*, 48(94):129–152, 2022. doi: 10.22210/suvlin.2022.094.01.

Pierre-André Bois. Hermann Schüttler, *Die Mitglieder des Illuminatenordens 1776–1787/93*, 1991. *Francia*, 20(2):248–249, 1993. doi: 10.11588/fr.1993.2.58307. URL https://journals.ub.uni-heidelberg.de/index.php/fr/article/view/58307. Rezension.

Rishi Bommasani et al. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Marta Boscariol, Luana Bulla, Lia Draetta, Beatrice Fiumanò, Emanuele Lenzi, and Leonardo Piano. Evaluation of LLMs on long-tail entity linking in historical documents, 2025. URL https://arxiv.org/abs/2505.03473.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, December 2011. doi: 10.1109/TVCG.2011.185. URL https://doi.org/10.1109/TVCG.2011.185.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. In Christian Chiarcos, Richard Eckart de Castilho, and Manfred Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten. Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Tübingen, 2009. Narr Francke Attempto Verlag. ISBN 9783823375111.

J. Scott Brennen and Daniel Kreiss. *Digitalization*, pages 1–11. John Wiley & Sons, Ltd, 2016. ISBN 9781118766804. doi: https://doi.org/10.1002/9781118766804.wbiect111. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118766804.wbiect111.

Alexander Brey. Digital art history in 2021. *History Compass*, 19(8):e12678, 2021. doi: https://doi.org/10.1111/hic3.12678. URL https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/hic3.12678.

Dan Brickley and R. V. Guha. RDF vocabulary description language 1.0: RDF schema. W3c recommendation, W3C, 2004. URL https://www.w3.org/TR/rdf-schema/.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners, 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Enrica Bruno, Maria Francesca Bocchi, and Francesca Tomasi. Modeling intermediality and interpretations in contemporary combinatory literature: Revealing Il Giuoco dell'Oca by Edoardo Sanguineti. In Simone Rebora, Marco Rospocher, and Stefano Bazzaco, editors, *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale – Proceedings del XIV Convegno Annuale AIUCD 2025*, Verona, Italy, 2025. AIUCD. ISBN 978-88-942535-9-7. doi: 10.6092/unibo/amsacta/8380. URL https://aiucd2025.dlls.univr.it/assets/pdf/papers/99.pdf.

Oleksandra Bruns, Tabea Tietz, Joerg Waitelonis, Etienne Posthumus, and Harald Sack. NFDIcore 2.0: A BFO-compliant ontology for multi-domain research infrastructures, 2024. URL https://arxiv.org/abs/2410.01821.

Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp. *Digital_Humanities*. The MIT Press, November 2012. ISBN 978-0-262-31210-3. doi: 10.7551/mitpress/9248.001.0001. URL https://doi.org/10.7551/mitpress/9248.001.0001.

Matteo Burioni. Giorgio vasaris vite. In Wolfgang Brassat, editor, *Handbuch Rhetorik der Bildenden Künste*, pages 269–282. De Gruyter, 2017. doi: 10.1515/9783110331493-015.

Peter Burke. *A Social History of Knowledge: From the Encyclopaedia to Wikipedia*, volume 2. Polity Press, Cambridge, 2012.

Lou Burnard. *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources*. OpenEdition Press, Marseille, 2014. doi: 10.4000/books.oep.426. URL http://books.openedition.org/oep/426.

R. Busa. The annals of humanities computing: The index thomisticus. *Computers and the Humanities*, 14(2):83–90, October 1980. ISSN 1572-8412. doi: 10.1007/BF02403798.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Navid Nobani, and Andrea Seveso. XAI meets LLMs: A Survey of the Relation between Explainable AI and Large Language Models, July 2024. URL http://arxiv.org/abs/2407.15248. doi: 10.48550/arXiv.2407.15248.

Stephen J. Campbell. *Vasari's Renaissance and Its Renaissance Alternatives*, pages 47–67. Routledge, 2008.

Rafael Capurro. Digital hermeneutics: an outline. *AI & SOCIETY*, 25(1):35–42, April 2010. ISSN 1435-5655. doi: 10.1007/s00146-009-0255-9.

Gavin Carothers and Andy Seaborne. RDF 1.1 N-Triples: A line-based syntax for an RDF graph. W3c recommendation, W3C, February 2014. URL https://www.w3.org/TR/n-triples/.

Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. ArCo: The Italian cultural heritage knowledge graph. In Chiara Ghidini et al., editors, *The Semantic Web – ISWC 2019*, volume 11779 of *Lecture Notes in Computer Science*, pages 36–52, Cham, 2019. Springer. doi: 10.1007/978-3-030-30796-7_3.

Valentina Anita Carriero, Marilena Daquino, Aldo Gangemi, Andrea Giovanni Nuzzolese, Silvio Peroni, Valentina Presutti, and Francesca Tomasi. *The Landscape of Ontology Reuse Approaches*, pages 21–38. IOS Press, 2020. doi: 10.3233/ssw200033. URL http://dx.doi.org/10.3233/SSW200033.

Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. Pattern-based design applied to cultural heritage knowledge graphs. *Semantic Web*, 12(2):313–339, 2021. doi: 10.3233/SW-200422. URL https://content.iospress.com/articles/semantic-web/sw200422.

Marc Cavazza and Pierre Zweigenbaum. Extracting implicit information from free text technical reports. *Information Processing & Management*, 28(5):609–618, September 1992. ISSN 0306-4573. doi: 10.1016/0306-4573(92)90030-4. URL https://www.sciencedirect.com/science/article/pii/0306457392900304.

Miquel Centelles Velilla and Núria Ferran-Ferrer. Using wikidata's ontology in practice: A neuro-symbolic, community-centred workflow for integrating and reusing humanities datasets. *Journal of Open Humanities Data*, December 2025. doi: 10.5334/johd.439.

William Charles. *From document and data interpretation to historical knowledge : an ontological approach applied to the case of territories*. Theses, Université de Toulouse, July 2025. URL https://theses.hal.science/tel-05262507.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6), June 2025a. ISSN 2666-3899. doi: 10.1016/j.patter.2025.101260. URL https://doi.org/10.1016/j.patter.2025.101260.

Jianhui Chen, Yuzhang Luo, and Liangming Pan. Mechanistic Data Attribution: Tracing the Training Origins of Interpretable LLM Units, January 2026. URL http://arxiv.org/abs/2601.21996. doi: 10.48550/arXiv.2601.21996.

Lihu Chen, Simon Razniewski, and Gerhard Weikum. Knowledge base completion for long-tail entities. In Estevam Hruschka, Tom Mitchell, Sajjadur Rahman, Dunja Mladenić, and Marko Grobelnik, editors, *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023)*, page 99–108, Toronto, ON, Canada, 2023. Association

for Computational Linguistics. doi: 10.18653/v1/2023.matching-1.8. URL https://aclanthology.org/2023.matching-1.8/.

Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, Xunliang Cai, Junxian He, and Jingang Wang. Revisiting scaling laws for language models: The role of data quality and training strategies. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23881–23899, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1163. URL https://aclanthology.org/2025.acl-long.1163/.

Nancy Chinchor. MUC-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference*, 1998.

Andrea Cigliano, Francesca Fallucchi, and Marco Gerardi. The impact of digital analysis and large language models in digital humanity. In *Proceedings of ICYRIME 2024*, pages 1–8, 2024. URL https://ceur-ws.org/Vol-3869/p01.pdf.

Ashley M. Clark and Steven W. Holloway. 'The past is never dead. It's not even past': The challenge of data provenance in the e-Humanities. In *Digital Humanities 2012: Conference Abstracts*, Hamburg, Germany, 2012. University of Hamburg. URL https://www-archiv.fdm.uni-hamburg.de/dh2012/conference/programme/abstracts/the-past-is-never-dead-its-not-even-past-the-challenge-of-data-provenance-in-the-e-humanities.1.html.

Alessio Cocchieri, Marcos Martínez Galindo, Giacomo Frisoni, Gianluca Moro, Claudio Sartori, and Giuseppe Tagliavini. ZeroNER: Fueling zero-shot named entity recognition via entity type descriptions. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15594–15616, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.805. URL https://aclanthology.org/2025.findings-acl.805/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Bruce Cole. *The Renaissance Artist at Work: From Pisano to Titian*. Routledge, 1 edition, 1983. ISBN 9780064301299.

Kate Crawford. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, New Haven, 2021.

Sarah Crowe, Kathrin Cresswell, Ann Robertson, Guro Huby, Anthony Avery, and Aziz Sheikh. The case study approach. *BMC Medical Research Methodology*, 11(1):100, 2011. doi: 10.1186/1471-2288-11-100.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, February 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45563-x.

Mathieu d'Aquin and Natalya F. Noy. Where to publish and find ontologies? a survey of ontology libraries. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:96–111, 2012. doi: 10.1016/j.websem.2011.08.005.

Marilena Daquino and Francesca Tomasi. Historical context ontology (HiCO): A conceptual model for describing context information of cultural heritage objects. In Emmanouel Garoufallou, Richard J. Hartley, and Panorea Gaitanou, editors, *Metadata and Semantics Research*, page 424–436, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24129-6. doi: https://doi.org/10.1007/978-3-319-24129-6_37.

Susan B. Davidson and Juliana Freire. Provenance and scientific workflows: Challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1345–1350, New York, NY, USA, 2008. ACM. doi: 10.1145/1376616.1376772.

DCMI Usage Board. DCMI metadata terms, January 2020. URL `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/`.

Jacopo de Berardinis, Valentina Anita Carriero, Nitisha Jain, Nicolas Lazzari, Albert Meroño-Peñuela, Andrea Poltronieri, and Valentina Presutti. The polifonia ontology network: Building a semantic backbone for musical heritage. In Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, editors, *The Semantic Web – ISWC 2023*, page 302–322, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-47243-5.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=5k8F6UU39V`.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290, 2022. doi: 10.1162/tacl_a_00460. URL https://aclanthology.org/2022.tacl-1.16/.

Ewa Deelman, Bruce Berriman, Ann Chervenak, Oscar Corcho, Paul Groth, and Luc Moreau. Metadata and provenance management. In Arie Shoshani and Doron Rotem, editors, *Scientific Data Management: Challenges, Technology, and Deployment*, pages 433–466. CRC Press, 2009. doi: 10.1201/9781420069815.

Shumin Deng, Ningyu Zhang, Luoqiu Li, Chen Hui, Tou Huaixiao, Mosha Chen, Fei Huang, and Huajun Chen. OntoED: Low-resource event detection with ontology embedding. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2828–2839, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.220. URL https://aclanthology.org/2021.acl-long.220/.

Guillaume Desagulier. *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Quantitative Methods in the Humanities and Social Sciences. Springer, Cham, 2017. doi: 10.1007/978-3-319-64572-8.

Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, and Enrico Motta. Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. *Future Generation Computer Systems*, 116: 253–264, March 2021. ISSN 0167-739X. URL https://publikationen.bibliothek.kit.edu/1000126629.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Jürgen Dietrich and André Hollstein. Performance and Reproducibility of Large Language Models in Named Entity Recognition: Considerations for the Use in Controlled Environments. *Drug Safety*, 48(3):287–303, March 2025. ISSN 1179-1942. doi: 10.1007/s40264-024-01499-1. URL https://doi.org/10.1007/s40264-024-01499-1.

Yifan Ding, Michael Yankoski, and Tim Weninger. Span-oriented information extraction: A unified framework. *SIGKDD Explor. Newsl.*, 27(1):137–155, July 2025. ISSN 1931-0145. doi: 10.1145/3748239.3748250. URL https://doi.org/10.1145/3748239.3748250.

James E. Dobson. The digitalization of the humanities: Recent trends and future prospects. In Mirko Farina, Paolo Ciancarini, Xiao Yu, and Jin Chen, editors, *Digital Transformation in Artificial Systems*, pages 313–329. Morgan Kaufmann, 2026. ISBN 978-0-443-32964-7. doi: https://doi.org/10.1016/B978-0-443-32964-7.00016-6. URL https://www.sciencedirect.com/science/article/pii/B9780443329647000166.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL https://aclanthology.org/L04-1011/.

Martin Doerr. The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75, 2003. doi: 10.1609/aimag.v24i3.1720.

Martin Doerr, Pavlos Fafalios, Achille Felicetti, Francesca Murano, Athanasios Velios, and Eleni Tsouloucha. Definition of the CRMtex: An extension of CIDOC CRM to model textual entities. Approved by the CIDOC CRM-SIG.

Martin Doerr, Athina Kritsotaki, and Katerina Boutsika. Factual argumentation—a core model for assertions making. *ACM Journal on Computing and Cultural Heritage*, 3(3):1–34, 2011. doi: 10.1145/1921614.1921615.

Martin Doerr, Stephen Stead, and Maria Theodoridou. CRMdig: An extension of CIDOC-CRM to support provenance metadata, version 3.2.1. Technical report, FORTH-ICS, April 2016. URL https://cidoc-crm.org/sites/default/files/CRMdig_v3.2.1.pdf.

Martin Doerr, Christian-Emil Ore, Pavlos Fafalios, Athina Kritsotaki, and Stephen Stead. Definition of the CRMinf: An extension of CIDOC-CRM to support argumentation, October 2023. URL https://cidoc-crm.org/sites/default/files/CRMinf%20v1.0(site).pdf.

Hang Dong, Jiaoyan Chen, Yuan He, Yinan Liu, and Ian Horrocks. Reveal the unknown: Out-of-knowledge-base mention discovery with entity linking. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 452–462. ACM, October 2023. doi: 10.1145/3583780.3615036. URL http://dx.doi.org/10.1145/3583780.3615036.

Kristina Dorey. Vasari's biography of Bazzi as 'Soddoma': Art history and literary analysis. *Italian Studies*, 70(2):167–190, 2015. doi: 10.1179/0075163415Z.00000000094.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Johann Gustav Droysen. *Grundriss der Historik*. Verlag von Veit & Comp., Leipzig, 1868.

Johanna Drucker. Humanities Approaches to Graphical Display. *DHQ: Digital Humanities Quarterly*, 5(1), 2011. URL https://dhq.digitalhumanities.org/vol/5/1/000091/000091.html.

Eamon Duede. Deep learning opacity in scientific discovery. *Philosophy of Science*, 90(5): 1089–1099, 2023. doi: 10.1017/psa.2023.8.

Brian D. Earp, Sebastian Porsdam Mann, et al. LLM use in scholarly writing poses a provenance problem. *Nature Machine Intelligence*, 7:1889–1890, 2025. doi: 10.1038/s42256-025-01159-8.

Carla Ebel, Jouni Tuominen, Matthias Schlögl, and Petri Leskinen. InTaVia/idm-rdf: IDM-RDF stable release v1.1, December 2023. URL https://doi.org/10.5281/zenodo.10290685. doi: 10.5281/zenodo.10290685.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association (CLEF 2020)*, volume 12260 of *Lecture Notes in Computer Science*, pages 288–310. Springer, 2020. doi: 10.1007/978-3-030-58219-7_21.

Maud Ehrmann, Matteo Romanello, Antoine Doucet, and Simon Clematide. Introducing the HIPE 2022 shared task: Named entity recognition and linking in multilingual historical documents. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval: 44th European Conference on IR Research (ECIR 2022)*, volume 13186 of *Lecture Notes in Computer Science*, pages 347–354. Springer, 2022. doi: 10.1007/978-3-030-99739-7_44.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM*

Comput. Surv., 56(2), September 2023. ISSN 0360-0300. doi: 10.1145/3604931. URL https://doi.org/10.1145/3604931.

Thomas Eiter, Giovambattista Ianni, Axel Polleres, Roman Schindlauer, and Hans Tompits. *Reasoning with Rules and Ontologies*, page 93–127. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-38412-0. doi: 10.1007/11837787_4. URL https://doi.org/10.1007/11837787_4.

Anthony Elliott. *The Culture of AI: Everyday Life and the Digital Revolution*. Routledge, London and New York, 2019. ISBN 978-1-138-23005-7. doi: 10.4324/9781315387185.

Orri Erling and Ivan Mikhailov. *RDF Support in the Virtuoso DBMS*, page 7–24. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-02184-8. doi: 10.1007/978-3-642-02184-8_2. URL https://doi.org/10.1007/978-3-642-02184-8_2.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, December 2008. ISSN 0001-0782, 1557-7317. doi: 10.1145/1409360.1409378. URL https://dl.acm.org/doi/10.1145/1409360.1409378.

Riccardo Falco, Aldo Gangemi, Silvio Peroni, David Shotton, and Fabio Vitali. Modelling OWL ontologies with graffoo. In Valentina Presutti, Eva Blomqvist, Raphaël Troncy, Harald Sack, Ioannis Papadakis, and Anna Tordai, editors, *The Semantic Web: ESWC 2014 Satellite Events*, volume 8798 of *Lecture Notes in Computer Science*, page 320–325, Cham, 2014. Springer. doi: 10.1007/978-3-319-11955-7_42.

Yuchen Fan, Yantao Liu, Zijun Yao, Jifan Yu, Lei Hou, and Juanzi Li. Evaluating Generative Language Models in Information Extraction as Subjective Question Correction, April 2024. URL http://arxiv.org/abs/2404.03532. doi: 10.48550/arXiv.2404.03532.

Reza Fayyazi, Stella Hoyos Trueba, Michael Zuzak, and Shanchieh Jay Yang. ProveRAG: Provenance-driven vulnerability analysis with automated retrieval-augmented LLMs. *IEEE Access*, 13:212815–212826, 2025. doi: 10.1109/ACCESS.2025.3638251.

Edward A. Feigenbaum. The art of artificial intelligence: Themes and case studies of knowledge engineering. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1014–1029, 1977.

Achille Felicetti and Francesca Murano. Ce qui est écrit et ce qui est parlé. crmtex for modelling textual entities on the semantic web. *Semant. Web*, 12(2):169–180, January 2021. ISSN 1570-0844. doi: 10.3233/SW-200418. URL https://doi.org/10.3233/SW-200418.

Mariano Fernández, Asunción Gómez-Pérez, and Natalia Juristo. METHONTOLOGY: From ontological art towards ontological engineering. In *Proceedings of the AAAI Spring Symposium on Ontological Engineering*, AAAI Technical Report SS-97-06, pages 33–40, Menlo Park, CA, 1997. AAAI Press. URL https://cdn.aaai.org/Symposia/Spring/1997/SS-97-06/SS97-06-005.pdf.

Andreas Fickers. Update für die hermeneutik. geschichtswissenschaft auf dem weg zur digitalen forensik? *Zeithistorische Forschungen / Studies in Contemporary History*, 17(1):157–168, 2020. doi: 10.14765/zzf.dok-1765.

Kerstin Fischer. Cognitive Linguistics and Pragmatics. In Barbara Dancygier, editor, *The Cambridge Handbook of Cognitive Linguistics*, Cambridge Handbooks in Language and Linguistics, pages 330–346. Cambridge University Press, Cambridge, 2017. ISBN 978-1-107-54420-8. doi: 10.1017/9781316339732.021. URL https://www.cambridge.org/core/product/396F85C58BF898ED2081120BD4E00580.

Julia Flanders and Fotis Jannidis, editors. *The Shape of Data in Digital Humanities: Modeling Texts and Text-based Resources*. Digital Research in the Arts and Humanities. Routledge, London and New York, 2019. ISBN 978-1-4724-4324-3.

Luciano Floridi. *The Ethics of Information*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-964132-1. doi: 10.1093/acprof:oso/9780199641321.001.0001.

Luciano Floridi and Massimo Chiriatti. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, December 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09548-1.

Antske Fokkens, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, and Victor de Boer. BiographyNet: Extracting relations between people and events, 2018. URL https://arxiv.org/abs/1801.07073.

Michel Foucault. *Power/Knowledge: Selected Interviews and Other Writings, 1972–1977*. Pantheon Books, New York, 1980.

Michel Foucault. *This Is Not a Pipe*. University of California Press, Berkeley, 1983. ISBN 978-0-520-04916-1.

Gottlob Frege. *Sense and Reference*, page 56–78. Blackwell, Oxford, 1948.

Yufeng Fu, Xiangge Li, Ce Shang, Hong Luo, and Yan Sun. ZoIE: A zero-shot open information extraction model based on language model. In *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 784–789, 2023. doi: 10.1109/CSCWD57460.2023.10152821.

Hans-Georg Gadamer. *Truth and Method*. Seabury Press, New York, 1960.

Selina Galka and Georg Vogeler. Annotating, projecting, and interpreting named entities in digital scholarly editions with LLMs. *International Journal of Digital Humanities*, 7(3): 483–509, December 2025. ISSN 2524-7840. doi: 10.1007/s42803-025-00114-8. URL https://doi.org/10.1007/s42803-025-00114-8.

Fabien Gandon and Guus Schreiber. RDF 1.1 XML syntax. W3c recommendation, W3C, February 2014. URL https://www.w3.org/TR/rdf-syntax-grammar/.

Aldo Gangemi and Valentina Presutti. *Ontology Design Patterns*, page 221–243. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-92673-3. doi: 10.1007/978-3-540-92673-3_10. URL https://doi.org/10.1007/978-3-540-92673-3_10.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. Sweetening WORDNET with DOLCE. *AI Magazine*, 24(3):13, 2003. doi: 10.1609/aimag.v24i3.1715.

Aldo Gangemi, Carola Catenacci, Massimiliano Ciaramita, and Jos Lehmann. Modelling Ontology Evaluation and Validation. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications (ESWC 2006)*, volume 4011 of *Lecture Notes in Computer Science*, pages 140–154, Berlin, Heidelberg, 2006. Springer. doi: 10.1007/11762256_13.

Aldo Gangemi, Arianna Graciotti, Antonello Meloni, Eleonora Marzi, Andrea Nuzzolese, Valentina Presutti, Diego Reforgiato Recupero, Alessandro Russo, and Rocco Tripodi. MusicBO: An application of Text2AMR2FRED to the musical heritage domain. In *Proceedings of the 21st Extended Semantic Web Conference (ESWC 2024) Satellite Events*, 2024. URL https://2024.eswc-conferences.org/wp-content/uploads/2024/05/77770206.pdf.

Aldo Gangemi, Andrea Graciotti, Eleonora Marzi, Andrea Meloni, Andrea Giovanni Nuzzolese, Valentina Presutti, Diego Reforgiato Recupero, Alessandro Russo, and Rocco Tripodi. Text2AMR2FRED: Converting text into RDF/OWL knowledge graphs via abstract meaning representation. *Knowledge and Information Systems*, 2025. doi: 10.1007/s10115-025-02631-y.

Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. In *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part II*, page 676–691, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-11011-6. doi: 10.1007/978-3-030-11012-3_52. URL https://doi.org/10.1007/978-3-030-11012-3_52.

Leyla Jael García Castro, Olga X. Giraldo, and Alexander García Castro. Using the annotation ontology in semantic digital libraries. In *Proceedings of the ISWC 2010 Posters&*

Demonstrations Track: Collected Abstracts, volume 658 of *CEUR Workshop Proceedings*, page 153–156, Shanghai, China, 2010. CEUR-WS.org. URL https://ceur-ws.org/Vol-658/paper505.pdf.

Darío Garigliotti, Dyaa Albakour, Miguel Martinez, and Krisztian Balog. Unsupervised context retrieval for long-tail entities. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 225–228, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6881-0. doi: 10.1145/3341981.3344244. URL https://doi.org/10.1145/3341981.3344244.

Daniel Garijo. WIDOCO: A wizard for documenting ontologies. In Claudia d'Amato, Miriam Fernandez, Valentina Tamma, Freddy Lecue, Philippe Cudré-Mauroux, Juan Sequeda, Christoph Lange, and Jeff Heflin, editors, *The Semantic Web – ISWC 2017*, page 94–102, Cham, 2017. Springer International Publishing. ISBN 978-3-319-68204-4.

Gemini Team, Google, Rohan Anil, et al. Gemini: A family of highly capable multimodal models. Technical report, Google, 2024. URL https://arxiv.org/abs/2312.11805.

Gemma Team, Google. Gemma 4 technical report. Technical report, Google, 2026.

Barbara Gendron, Gaël Guibon, and Mathieu d'Aquin. Conversational control with ontologies for large language models: A lightweight framework for constrained generation, 2026. URL https://arxiv.org/abs/2604.04450.

Genet Asefa Gesese, Zongxiong Chen, Shufan Jiang, Mary Ann Tan, Zhaotai Liu, Sonja Schimmler, and Harald Sack. NERdME: A named entity recognition dataset for indexing research artifacts in code repositories. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 8661–8664, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 979-8-4007-2307-0. doi: 10.1145/3774904.3792934. URL https://doi.org/10.1145/3774904.3792934.

Kerstin Gierend, Frank Krüger, Sascha Genehr, Francisca Hartmann, Fabian Siegel, Dagmar Waltemath, Thomas Ganslandt, and Atinkut Alamirrew Zeleke. Provenance information for biomedical data and workflows: Scoping review. *Journal of Medical Internet Research*, 26: e51297, 2024. doi: 10.2196/51297.

Carlo Ginzburg, John Tedeschi, Anne C. Tedeschi, and Carlo Ginzburg. *Clues, Myths, and the Historical Method*. Johns Hopkins University Press, 2013. ISBN 978-1-4214-0990-0 978-1-4214-0991-7. doi: 10.56021/9781421409900. URL https://www.press.jhu.edu/books/title/10644/clues-myths-and-historical-method.

John Giorgi, Gary D. Bader, and Bo Wang. A sequence-to-sequence approach for document-level relation extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 10–25, 2022. doi: 10.18653/v1/2022.bionlp-1.2. URL https://aclanthology.org/2022.bionlp-1.2/.

Birte Glimm, Ian Horrocks, Boris Motik, Giorgos Stoilos, and Zhe Wang. HermiT: an OWL 2 reasoner. *Journal of automated reasoning*, 53(3):245–269, 2014.

Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. Collective entity resolution with multi-focal attention. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1059. URL https://aclanthology.org/P16-1059/.

Arianna Graciotti, Leonardo Piano, Nicolas Lazzari, Enrico Daga, Rocco Tripodi, Valentina Presutti, and Livio Pompianu. KE-MHISTO: Towards a Multilingual Historical Knowledge Extraction Benchmark for Addressing the Long-Tail Problem. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20316–20339, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1042. URL https://aclanthology.org/2025.findings-acl.1042/.

Stefan Gradmann and Jan Christoph Meister. Digital document and interpretation: Re-thinking "text" and scholarship in electronic settings. *Poiesis & Praxis*, 5(2):139–153, 2008. doi: 10.1007/s10202-007-0042-y.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. doi: 10.48550/arXiv.2407.21783.

Ralph Grishman. Information extraction: Techniques and challenges. In Maria Teresa Pazienza, editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, pages 10–27. Springer, 1997. doi: 10.1007/3-540-63438-X_2.

Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993. doi: 10.1006/knac.1993.1008.

Michael Grüninger and Mark S. Fox. Methodology for the design and evaluation of ontologies. In *Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995. URL https://www.cs.toronto.edu/kr/publications/.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: On-policy distillation of large language models, 2026. URL https://arxiv.org/abs/2306.08543.

Nicola Guarino. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, 43(5):625–640, November 1995. ISSN 1071-5819. doi: 10.1006/ijhc.1995.1066.

Nicola Guarino. Understanding, building and using ontologies. *International Journal of Human-Computer Studies*, 46(2):293–310, February 1997. ISSN 1071-5819. doi: 10.1006/ijhc.1996.0091.

Nicola Guarino and Pierdaniele Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. In N. J. I. Mars, editor, *Towards Very Large Knowledge Bases*, page 25–32, Amsterdam, 1995. IOS Press.

Nicola Guarino and Christopher Welty. Evaluating ontological decisions with OntoClean. *Commun. ACM*, 45(2):61–65, February 2002. ISSN 0001-0782. doi: 10.1145/503124.503150. URL https://doi.org/10.1145/503124.503150.

Nicola Guarino, Daniel Oberle, and Steffen Staab. *What Is an Ontology?*, page 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-92673-3. doi: 10.1007/978-3-540-92673-3_0. URL https://doi.org/10.1007/978-3-540-92673-3_0.

Takuo Hamaguchi, Hidekazu Oiwa, Masashi Shimbo, and Yuji Matsumoto. Knowledge Transfer for Out-of-Knowledge-Base Entities: A Graph Neural Network Approach. *Transactions of the Japanese Society for Artificial Intelligence*, 33(2):F–H72_1–10, 2018. ISSN 1346-0714, 1346-8030. doi: 10.1527/tjsai.F-H72. URL http://arxiv.org/abs/1706.05674.

Ridong Han, Chaohao Yang, Tao Peng, Prayag Tiwari, Xiang Wan, Lu Liu, and Benyou Wang. An Empirical Study on Information Extraction using Large Language Models, September 2024. URL http://arxiv.org/abs/2305.14450. doi: 10.48550/arXiv.2305.14450.

Muhammad Uzair Ul Haq, Davide Rigoni, and Alessandro Sperduti. LLMs as Data Annotators: How Close Are We to Human Performance, April 2025. URL http://arxiv.org/abs/2504.15022. doi: 10.48550/arXiv.2504.15022.

Patricia Harpring. Development of the Getty Vocabularies: AAT, TGN, ULAN, and CONA. *Art Documentation: Journal of the Art Libraries Society of North America*, 29(1):67–72, 2010. doi: 10.1086/adx.29.1.27949541.

Steve Harris and Andy Seaborne. SPARQL 1.1 query language. W3c recommendation, W3C, 2013. URL https://www.w3.org/TR/sparql11-query/.

Olaf Hartig. Foundations of RDF* and SPARQL*: An alternative approach to statement-level metadata in RDF. In *Proceedings of the AMW 2017 Workshop on Alberto Mendelzon International Workshop on Foundations of Data Management*, volume 1912 of *CEUR Workshop Proceedings*, 2017. URL https://ceur-ws.org/Vol-1912/paper12.pdf.

Hashmy Hassan, Sudheep Elayidom, M. R. Irshad, and Christophe Chesneau. A detailed analysis into evaluation metrics for knowledge graph evaluation. *International Journal of Data Science and Analytics*, 20:6933–6972, 2025. doi: 10.1007/s41060-025-00871-3. URL https://doi.org/10.1007/s41060-025-00871-3.

Patrick J. Hayes and Peter F. Patel-Schneider. RDF 1.1 semantics. W3c recommendation, W3C, 2014. URL https://www.w3.org/TR/rdf11-mt/.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38. Association for Computational Linguistics, 2010.

Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Q.*, 28(1):75–105, March 2004. ISSN 0276-7783.

Pascal Hitzler, Markus Krötzsch, and Sebastian Rudolph. *Foundations of Semantic Web Technologies*. CRC Press, 2010. ISBN 9781420090505.

Pascal Hitzler, Monireh Ebrahimi, Md Kamruzzaman Sarker, and Daria Stepanova. Neuro-symbolic AI and the semantic web. *Semantic Web*, 15(5):1201–1207, 2024. doi: 10.3233/SW-243711. URL https://journals.sagepub.com/doi/10.3233/SW-243711.

Hlomani Hlomani and Deborah Stacey. Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *Semantic Web Journal*, 1(5):1–11, 2014.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Computing Surveys*, 54(4):71:1–71:37, 2021. doi: 10.1145/3447772.

Richard Hohensinner, Belgin Mutlu, Inti Gabriel Mendoza Estrada, Matej Vukovic, Simone Kopeinik, and Roman Kern. Tracing the Data Trail: A Survey of Data Provenance, Transparency and Traceability in LLMs, January 2026. URL http://arxiv.org/abs/2601.14311. doi: 10.48550/arXiv.2601.14311.

Hawre Hosseini. Implicit Entity Recognition, Classification and Linking in Tweets. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1448, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6172-9. doi: 10.1145/3331184.3331416. URL https://doi.org/10.1145/3331184.3331416.

Hawre Hosseini and Ebrahim Bagheri. Learning to rank implicit entities on twitter. *Information Processing& Management*, 58(3):102503, May 2021. ISSN 0306-4573. doi: 10.1016/j.ipm.2021.102503.

Hawre Hosseini, Tam T. Nguyen, and Ebrahim Bagheri. Implicit entity linking through ad-hoc retrieval. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 326–329, 2018. doi: 10.1109/ASONAM.2018.8508612.

Peyman Hosseini, Ignacio Castro, Iacopo Ghinassi, and Matthew Purver. Efficient Solutions For An Intriguing Failure of LLMs: Long Context Window Does Not Mean LLMs Can Analyze Long Sequences Flawlessly. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1880–1891, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.128/.

Eduard Hovy and Julia Lavid. Towards a science of corpus annotation. *International Journal of Translation*, 22(1):1–23, 2010.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820, September 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocad259. URL https://doi.org/10.1093/jamia/ocad259.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2):1–55, March 2025. ISSN 1046-8188, 1558-2868. doi: 10.1145/3703155. URL http://arxiv.org/abs/2311.05232.

Pere-Lluis Huguet-Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. RED^{fm}: a filtered and multilingual relation extraction dataset. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4326–4343, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.237. URL https://aclanthology.org/2023.acl-long.237/.

Eero Hyvönen. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2012. doi: 10.2200/S00452ED1V01Y201210WBE003.

Eero Hyvönen. Digital humanities on the semantic web: Sampo model and portal series. *Semantic Web*, 14(4):729–744, 2023. ISSN 2210-4968. doi: 10.3233/SW-223034.

Eero Hyvönen. Digital humanities on the semantic web: from infrastructure to practical applications, AI-Based knowledge discovery, and web of wisdom. In Wolf-Tilo Balke, Koraljka Golub, Yannis Manolopoulos, Kostas Stefanidis, and Zheying Zhang, editors, *Linking Theory and Practice of Digital Libraries*, page 3–8, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-05409-8.

IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional Requirements for Bibliographic Records: Final Report, 2009. URL https://repository.ifla.org/handle/20.500.14598/811.

Esko Ikkala, Eero Hyvönen, Heikki Rantala, and Mikko Koho. Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. *Semantic Web*, 13(1):69–84, 2022. doi: 10.3233/SW-210428.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing (STOC '98)*, pages 604–613, 1998. doi: 10.1145/276698.276876.

Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha, and Nurfadhlina Mohd Sharef. An analysis of ontology engineering methodologies: A literature review. *Research Journal of Applied Sciences, Engineering and Technology*, 6(16):2993–3000, 2013. doi: 10.19026/rjaset.6.3684.

Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1158. Association for Computational Linguistics, 2011.

Shengjie Ji and Fang Kong. A Novel Three-stage Framework for Few-shot Named Entity Recognition. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1293–1305, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.116/.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825. doi: 10.48550/arXiv.2310.06825.

Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. GenRES: Rethinking evaluation for generative relation extraction in the era of large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 2820–2837, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.155. URL https://aclanthology.org/2024.naacl-long.155/.

Xuhui Jiang, Chengjin Xu, Yinghan Shen, Xun Sun, Lumingyuan Tang, Saizhuo Wang, Zhongwu Chen, Yuanzhuo Wang, and Jian Guo. On the evolution of knowledge graphs: A survey and perspective, 2025. URL https://arxiv.org/abs/2310.04835.

Matthew L. Jockers. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, Urbana, 2013.

Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. GenIE: Generative Information Extraction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4626–4643, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.342. URL https://aclanthology.org/2022.naacl-main.342/.

Jürgen Jost and Apollonia Jost. Network analysis for the digital humanities: Principles, problems, extensions. *Isis*, 110(3):538–554, 2019. doi: 10.1086/705532.

Plinio Thomaz Aquino Junior and Lucia Vilela Leite Filgueiras. User modeling with personas. In *Proceedings of the 2005 Latin American Conference on Human-Computer Interaction*, CLIHC '05, page 277–282, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1-59593-224-0. doi: 10.1145/1111360.1111388. URL https://doi.org/10.1145/1111360.1111388.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition, 2026. URL https://web.stanford.edu/~jurafsky/slp3/.

Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In R. Borgo, F. Ganovelli, and I. Viola, editors, *Eurographics Conference on Visualization (EuroVis) – State of The Art Reports*. The Eurographics Association, 2015.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Deep Ganguli, Jackson Kernion, Andy Jones, Liane Lovitt, Zac Hatfield-Dodds, Dario Chen, Danny Hernandez, Dawn Drain, Nathaniel Landau, Kamal Ndousse, Amanda Shoker, Ethan Perez, Jack Clark, Jared Kaplan, Miles Brundage, Dario Amodei, Jack Clark, Sam McCandlish, Chris Olah, Nelson Elhage, Neel Nanda, William Saunders, Sandipan Kundu, et al. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Pankush Kalgotra and Ramesh Sharda. Network analytics: an introduction and illustrative applications in health data science. *Journal of Information Technology Case and Application Research*, 25(3):305–315, 2023. doi: 10.1080/15228053.2023.2187995. URL https://doi.org/10.1080/15228053.2023.2187995.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large Language Models Struggle to Learn Long-Tail Knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15696–15707. PMLR, July 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Andres Karjus. Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence. *Humanities and Social Sciences Communications*, 12

(1):277, February 2025. ISSN 2662-9992. doi: 10.1057/s41599-025-04503-w. URL https://www.nature.com/articles/s41599-025-04503-w.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondřej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondřej Dušek, and Simone Balloccu. LLMs as Span Annotators: A Comparative Study of LLMs and Humans, February 2026. URL http://arxiv.org/abs/2504.08697. doi: 10.48550/arXiv.2504.08697.

Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. Epistemic injustice in generative AI. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, page 684–697. AAAI Press, 2025.

C Maria Keet. *An Introduction to Ontology Engineering*. University of Cape Town, https://people.cs.uct.ac.za/ mkeet/files/OEbook.pdf, 2018. URL http://hdl.handle.net/11427/28312.

Natthawut Kertkeidkachorn and Ryutaro Ichise. T2KG: An end-to-end system for creating knowledge graphs from unstructured text. In *Proceedings of the AAAI Workshop on Knowledge Capture (K-CAP)*, 2017. URL https://cdn.aaai.org/ocs/ws/ws0328/15129-68427-1-PB.pdf.

Samira Khorshidi, Azadeh Nikfarjam, Suprita Shankar, Yisi Sang, Yash Govind, Hyun Jang, Ali Kasgari, Alexis McClimans, Mohamed Soliman, Vishnu Konda, Ahmed Fakhry, and Xiaoguang Qi. ODKE+: Ontology-Guided Open-Domain Knowledge Extraction with LLMs, September 2025. URL http://arxiv.org/abs/2509.04696. doi: 10.48550/arXiv.2509.04696.

Yeonghyeon Kim, Jungwoo Kim, Yunjin Jang, Mogan Lee, and Sungjoon Park. Evaluating span extraction in generative paradigm: A reflection on aspect-based sentiment analysis, 2024. URL https://arxiv.org/abs/2404.11539.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, 2018.

Georg Kloß. *Annalen der Loge zur Einigkeit, der Englischen Provincial-Loge, so wie der Provincial- und Directorial-Loge des eclectischen Bundes zu Frankfurt am Main: 1742–1811. Eine Festgabe, ausgetheilt bei der Säcularfeier der Loge zur Einigkeit am 27.\ Juni 1842*. Horstmann, Frankfurt am Main, 1842. URL https://www.digitale-sammlungen.de/de/details/bsb10446364.

Graham Klyne and Jeremy J. Carroll. Resource Description Framework (RDF): Concepts and abstract syntax. W3c recommendation, W3C, 2004. URL https://www.w3.org/TR/rdf-concepts/.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Nikolaos Kolitsas, Octavian-Emil Ganea, and Thomas Hofmann. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529. Association for Computational Linguistics, 2018.

Reinhart Koselleck. *Kritik und Krise. Eine Studie zur Pathogenese der bürgerlichen Welt*. Verlag Karl Alber, Freiburg and München, 1959.

Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. The open archives initiative protocol for metadata harvesting, version 2.0. Technical report, Open Archives Initiative, 2002. URL https://www.openarchives.org/OAI/openarchivesprotocol.html.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

Kusum Lata, Pardeep Singh, and Kamlesh Dutta. Mention detection in coreference resolution: survey. *Applied Intelligence*, 52(9):9816–9860, 2022. ISSN 1573-7497. doi: 10.1007/s10489-021-02878-2.

Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. PROV-O: The PROV ontology. W3C recommendation, World Wide Web Consortium, April 2013. URL https://www.w3.org/TR/prov-o/.

Melissa Geiger Leda Cempellin. *Reimagining Art History with Artificial Intelligence: Navigating Limitations and Pioneering New Paths for Innovation*. Palgrave Macmillan, Cham, 2026. ISBN 978-3-032-21684-7.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 188–197. Association for Computational Linguistics, 2017.

Ilya Levin. Epistemology of generative AI: The geometry of knowing, 2026. URL https://arxiv.org/abs/2602.17116.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A Survey on Fairness in Large Language Models, February 2024. URL http://arxiv.org/abs/2308.10149. doi: 10.48550/arXiv.2308.10149.

Xincheng Liao, Junwen Duan, Yixi Huang, and Jianxin Wang. RUIE: Retrieval-based Unified Information Extraction using Large Language Model. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9640–9655, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.645/.

Library of Congress. MARC XML schema. Technical report, Library of Congress Network Development and MARC Standards Office, 2002. URL https://www.loc.gov/standards/marcxml/.

Pasquale Lisena, Marieke van Erp, Cecilia Bembibre, and Inger Leemans. Data Mining and Knowledge Graphs as a Backbone for Advanced Olfactory Experiences. In *Smell, Taste, and Temperature Interfaces (STT 2021), workshop at CHI 2021*, Yokohama, Japan (Virtual), May 2021. ACM.

Pasquale Lisena, Daniel Schwabe, Marieke van Erp, Raphaël Troncy, William Tullett, Inger Leemans, Lizzie Marx, and Sofia Colette Ehrich. Capturing the semantics of smell: The odeuropa data model for olfactory heritage information. In Paul Groth, Maria-Esther Vidal, Fabian Suchanek, Pedro Szekley, Pavan Kapanipathi, Catia Pesquita, Hala Skaf-Molli, and Minna Tamper, editors, *The Semantic Web*, page 387–405, Cham, 2022. Springer International Publishing. ISBN 978-3-031-06981-9.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What Makes Good In-Context Examples for GPT-3?, January 2021. URL http://arxiv.org/abs/2101.06804. doi: 10.48550/arXiv.2101.06804.

Jiayu Liu, Qing Zong, Weiqi Wang, and Yangqiu Song. Revisiting epistemic markers in confidence estimation: Can markers accurately reflect large language models' uncertainty? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2:*

Short Papers), pages 206–221, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.18. URL https://aclanthology.org/2025.acl-short.18/.

Kaiyuan Liu, Shaotian Yan, Rui Miao, Bing Wang, Chen Shen, Jun Zhang, and Jieping Ye. Where did this sentence come from? tracing provenance in LLM reasoning distillation, 2026. URL https://arxiv.org/abs/2512.20908.

Lihui Liu and Kai Shu. Unifying knowledge in agentic LLMs: Concepts, methods, and recent advancements. *SIGKDD Explor. Newsl.*, 27(2):88–96, December 2026. ISSN 1931-0145. doi: 10.1145/3787470.3787478.

Pai Liu, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Zongsheng Li, Ehsan Hoque, Julia Hirschberg, and Yue Zhang. A survey on open information extraction from rule-based model to large language model, 2024. URL https://arxiv.org/abs/2208.08690.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), 2023a. ISSN 0360-0300. doi: 10.1145/3560815. URL https://doi.org/10.1145/3560815.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. Zero-shot entity linking by reading entity descriptions. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1335. URL https://aclanthology.org/P19-1335/.

Steffen Lohmann, Stefan Negru, Florian Haag, and Thomas Ertl. Visualizing ontologies with VOWL. *Semantic Web*, 7(4):399–419, May 2016. ISSN 1570-0844. doi: 10.3233/SW-150200.

Haoyan Luo and Lucia Specia. From understanding to utilization: A survey on explainability for large language models. *arXiv preprint arXiv:2401.12874*, 2024.

Thanh Luong Tuan. Ontology-constrained neural reasoning in enterprise agentic systems: A neurosymbolic architecture for domain-grounded AI agents. arXiv preprint arXiv:2604.00555, April 2026. URL https://arxiv.org/abs/2604.00555.

Steven Mailloux. *Interpretive Conventions: The Reader in the Study of American Fiction*. Cornell University Press, Ithaca, NY, 1982. ISBN 978-0-8014-1476-3. doi: 10.7591/9781501720949.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999. ISBN 9780262133609.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.645. URL `https://aclanthology.org/2020.acl-main.645/`.

Nestor Maslej et al. The AI index 2025 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2025. URL `https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf`.

Arcangelo Massari, Silvio Peroni, Francesca Tomasi, and Ivan Heibi. Representing provenance and track changes of cultural heritage metadata in RDF: A survey of existing approaches. *Digital Scholarship in the Humanities*, 41(Supplement_1):i196–i212, April 2026. doi: 10.1093/llc/fqaf076.

Willard McCarty. Towards a hermeneutics of scientific visualisation. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*. Blackwell, Oxford, 2004.

Willard McCarty. *Humanities Computing*. Palgrave Macmillan, London, 2005.

Deborah L. McGuinness and Frank van Harmelen. OWL web ontology language overview. W3c recommendation, W3C, 2004. URL `https://www.w3.org/TR/owl-features/`.

Igor Melnyk, Pierre Dognin, and Payel Das. Knowledge graph generation from text. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1610–1622, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.116. URL `https://aclanthology.org/2022.findings-emnlp.116/`.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

Albert Meroño-Peñuela et al. Semantic technologies for historical research: A survey. *Semantic Web*, 6(6), 2014. doi: 10.3233/SW-140158.

Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. Arctic-embed 2.0: Multilingual retrieval without compromise, 2024. URL https://arxiv.org/abs/2412.04506. doi: 10.48550/arXiv.2412.04506.

George Mihaila, Suleyman Olcay Polat, Poli Nemkova, Himanshu Sharma, Namratha V. Urs, and Mark V. Albert. LIME-LLM: Probing models with fluent counterfactuals, not broken text, 2026. URL https://arxiv.org/abs/2601.11746.

Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F. Enguix, and Kusum Lata. Text2KGBench: A benchmark for ontology-driven knowledge graph generation from text. In Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, editors, *The Semantic Web – ISWC 2023*, page 247–265, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-47243-5.

Tomasz Miksa, Simon Oblasser, and Andreas Rauber. Automating research data management using machine-actionable data management plans. *ACM Trans. Manage. Inf. Syst.*, 13(2), December 2021. ISSN 2158-656X. doi: 10.1145/3490396. URL https://doi.org/10.1145/3490396.

David Milne and Ian H. Witten. Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 509–518, 2008. doi: 10.1145/1458082.1458150.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large Language Models: A Survey, March 2025. URL http://arxiv.org/abs/2402.06196. doi: 10.48550/arXiv.2402.06196.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997. ISBN 978-0-07-042807-2.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681. URL https://aclanthology.org/2024.tacl-1.52/.

Christoph Molnar. *Interpretable Machine Learning*. Lean Publishing, 3 edition, 2025. ISBN 978-3-911578-03-5. URL https://christophm.github.io/interpretable-ml-book.

Franco Moretti. *Distant Reading*. Verso Books, 2013. ISBN 978-1-78168-333-0.

Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, Zhe Wu, Achille Fokoue, and Carsten Lutz. OWL 2 web ontology language profiles (second edition). W3c recommendation, W3C, 2012. URL https://www.w3.org/TR/owl2-profiles/.

Jennifer Mugleston, Vuong H. Truong, Cindy Kuang, Lungile Sibiya, and Jihwan Myung. Epistemology in the age of large language models. *Knowledge*, 5(1):3, 2025. ISSN 2673-9585. doi: 10.3390/knowledge5010003.

Tamara Munzner. *Visualization Analysis & Design*. AK Peters Visualization Series. CRC Press, 2014. ISBN 978-1466508910.

Peter Murray and Linda Murray. *The Art of the Renaissance*. World of Art. Thames & Hudson, London, 1963.

Cedric Möller and Ricardo Usbeck. Entity linking with out-of-knowledge-graph entity detection and clustering using only knowledge graphs. In Angelo A. Salatino, Mehwish Alam, Femke Ongenae, Sahar Vahdati, Anna Lisa Gentile, Tassilo Pellegrini, and Shufan Jiang, editors, *Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI – Proceedings of the 20th International Conference on Semantic Systems (SEMANTiCS 2024)*, volume 60 of *Studies on the Semantic Web*, pages 88–105, Amsterdam, The Netherlands, 2024. IOS Press. ISBN 978-1-64368-537-3. doi: 10.3233/SSW240009.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26, January 2007. ISSN 0378-4169, 1569-9927. doi: 10.1075/li.30.1.03nad. URL https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad.

Ebrahim Norouzi, Jörg Waitelonis, and Harald Sack. The landscape of ontologies in materials science and engineering: A survey and evaluation, 2024. URL https://arxiv.org/abs/2408.06034.

Alessandro Nova. Edition Giorgio Vasari, 2004. URL https://www.wagenbach.de/buecher/edition-giorgio-vasari.html.

Bernd Nowack. The semantic web: Not a piece of cake. Blog post, 2009. URL https://web.archive.org/web/20220628120341/http://bnode.org/blog/2009/07/08/the-semantic-web-not-a-piece-of-cake.

Christopher Nygren and Sonja Drimmer. Art History and AI: Ten Axioms. *International Journal for Digital Art History*, 9:5.02–5.13, April 2023. ISSN 2363-5401. doi: 10.11588/dah.2023.9.90400. URL https://ahnp.ub.uni-heidelberg.de/journals/dah/article/view/90400.

Claus Oberhauser. *Die verschwörungstheoretische Trias: Barruel – Robison – Starck.* Quellen und Darstellungen zur europäischen Freimaurerei. StudienVerlag, Innsbruck and Wien and Bozen, 2013.

Dominic Oldman, Martin Doerr, and Stefan Gradmann. Zen and the art of linked data: New strategies for a semantic web of humanist knowledge. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A New Companion to Digital Humanities*, pages 251–273. John Wiley & Sons, Hoboken, NJ, 2016. doi: 10.1002/9781118680605.ch18.

Alex Olieman, Kaspar Beelen, Milan van Lange, Jaap Kamps, and Maarten Marx. Good applications for crummy entity linkers? the case of corpus selection in digital humanities. In *Proceedings of the 13th International Conference on Semantic Systems (SEMANTiCS 2017)*, page 81–88, Amsterdam, Netherlands, 2017. doi: 10.1145/3132218.3132237.

Sarah Rebecca Ondraszek. Viewsari: Enabling New Perspectives on the Renaissance with a Knowledge Graph of Giorgio Vasari's *The* Lives. In *Proceedings of the ISWC 2025 Companion Volume*, volume 4085 of *CEUR-WS*, Nara, Japan, November 2025. CEUR-WS.org. URL `https://ceur-ws.org/Vol-4085/paper17.pdf`.

Sarah Rebecca Ondraszek, Grischka Petri, Ulrike Blumenthal, Lisa Dieckmann, Etienne Posthumus, and Harald Sack. eXtreme Design for Ontological Engineering in the Digital Humanities with Viewsari, a Knowledge Graph of Giorgio Vasari's *The* Lives. In *Proceedings of the 1st International Workshop of Semantic Digital Humanities (SemDH 2024)*, volume 3724 of *CEUR-WS*, Hersonissos, Greece, May 2024a. CEUR-WS.org. URL `https://ceur-ws.org/Vol-3724/paper5.pdf`.

Sarah Rebecca Ondraszek, Harald Sack, and Etienne Posthumus. Viewsari: New Perspectives on Historical Network Analysis in Giorgio Vasari's *The* Lives Using Knowledge Graphs. In Martin Grandjean, editor, *Historical Network Research Conference 2024 — Book of Abstracts*, pages 137–139, Lausanne, 2024b. Zenodo. doi: 10.5281/zenodo.12665443.

Sarah Rebecca Ondraszek, Bruno Sartini, Marieke van Erp, Andrea Poltronieri, Pasquale Lisena, and Harald Sack. One pattern to express them all? Towards Generalized Patterns for Ontology Design in the Digital Humanities. In *Proceedings of Workshop on Ontology Design and Patterns (WOP 2024)*, Baltimore, MD, United States of America, November 2024c. Zenodo. doi: 10.5281/zenodo.14063917. URL `https://www.fiz-karlsruhe.de/sites/default/files/FIZ/Dokumente/Forschung/ISE/Publications/Conferences-Workshops/WOP%40ISWC2024_OnePattern_ondraszek.pdf`.

Sarah Rebecca Ondraszek, Tabea Tietz, Etienne Posthumus, Heike Fliegl, Jörg Waitelonis, and Harald Sack. Indexing Historical Research Data: MemO and the NFDI4Memory Knowledge

Graph. In *Proceedings of the Conference on Research Data Infrastructure (CoRDI 2025)*, Aachen, DEU, August 2025a. doi: 10.5281/zenodo.16736125.

Sarah Rebecca Ondraszek, Tabea Tietz, Jörg Waitelonis, and Harald Sack. Interlinking Research Data and Services in the Historical Sciences with MemO and the NFDI4Memory Knowledge Graph. In *Proceedings of the 5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment*, Nara, Japan, November 2025b. CEUR Workshop Proceedings (CEUR-WS.org). doi: 10.5281/zenodo.17376359. URL https://ceur-ws.org/Vol-4065/paper09.pdf.

Sarah Rebecca Ondraszek, Jörg Waitelonis, Katja Keller, Claudia Niessner, Anna M. Jacyszyn, and Harald Sack. Ontologies in Motion: A BFO-Based Approach to Knowledge Graph Construction for Motor Performance Research Data in Sports Science. In *Proceedings of the 5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment*, volume Vol-4065 of *CEUR-WS*, pages 97–106, Nara, Japan, November 2025c. CEUR Workshop Proceedings. URL https://ceur-ws.org/Vol-4065/paper08.pdf.

Ontotext AD. GraphDB: Semantic graph database, 2024. URL https://www.ontotext.com/products/graphdb/.

OpenAI. ChatGPT: Optimizing language models for dialogue. Technical report, OpenAI, 2023. URL https://openai.com/blog/chatgpt.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, July 2024.

OpenAI. gpt-oss-120b and gpt-oss-20b model card. Technical report, OpenAI, 2025. URL https://openai.com/index/gpt-oss-model-card/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. GPT-4 technical report. Technical report, OpenAI, 2023. URL https://arxiv.org/abs/2303.08774.

Thomas Orgel, Martin Höffernig, Werner Bailer, and Silvia Russegger. A metadata model and mapping approach for facilitating access to heterogeneous cultural heritage assets. *International Journal on Digital Libraries*, 15(2):189–207, April 2015. ISSN 1432-1300. doi: 10.1007/s00799-015-0138-2.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. F-coref: Fast, accurate and easy to use coreference resolution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 48–56, Taipei, Taiwan, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-demo.6.

J. Neil Otte, John Beverley, and Alan Ruttenberg. BFO: Basic formal ontology. *Applied Ontology*, 17(1):17–43, March 2022. ISSN 1570-5838. doi: 10.3233/AO-220262.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Gabriele Padovani, Valentine Anantharaj, and Sandro Fiore. Provenance tracking in large-scale machine learning systems. In *Workshop Proceedings of the 54th International Conference on Parallel Processing*, 2025. doi: 10.1145/3750720.3757295.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. on Knowl. and Data Eng.*, 36(7):3580–3599, July 2024. ISSN 1041-4347. doi: 10.1109/TKDE.2024.3352100. URL https://doi.org/10.1109/TKDE.2024.3352100.

Sarah Parisot, Pedro M. Esperanca, Steven McDonagh, Tamas J. Madarasz, Yongxin Yang, and Zhenguo Li. Long-tail recognition via compositional knowledge transfer, 2022. URL https://arxiv.org/abs/2112.06741.

Ludovica Paseri and Massimo Durante. Examining epistemological challenges of large language models in law. *Cambridge Forum on AI: Law and Governance*, 1:e7, 2025. doi: 10.1017/cfl.2024.7.

Valentina Pasqual, Andrea Schimmenti, Francesca Tomasi, and Fabio Vitali. SEBI: Scholarly Evidence-Based Interpretation Ontology. Ontology specification, June 2024. URL https://valentinapasqual.github.io/sebi/.

Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508, December 2016. ISSN 22104968, 15700844. doi: 10.3233/SW-160218. URL https://journals.sagepub.com/doi/full/10.3233/SW-160218.

Sachin Pawar, Girish K. Palshikar, and Pushpak Bhattacharyya. Relation extraction: A survey. *arXiv preprint arXiv:1712.05191*, 2017.

Pathirage Perera. Knowledge-Driven Implicit Information Extraction. *Browse all Theses and Dissertations*, January 2016. URL https://corescholar.libraries.wright.edu/etd_all/1571.

Sujan Perera, Pablo Mendes, Amit Sheth, Krishnaprasad Thirunarayan, Adarsh Alex, Christopher Heid, and Greg Mott. Implicit entity recognition in clinical documents. In Martha

Palmer, Gemma Boleda, and Paolo Rosso, editors, *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 228–238, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-1028. URL https://aclanthology.org/S15-1028/.

Sujan Perera, Pablo N. Mendes, Adarsh Alex, Amit P. Sheth, and Krishnaprasad Thirunarayan. Implicit entity linking in tweets. In *Proceedings of the 13th International Conference on The Semantic Web. Latest Advances and New Domains - Volume 9678*, page 118–132, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 9783319341286. doi: 10.1007/978-3-319-34129-3_8. URL https://doi.org/10.1007/978-3-319-34129-3_8.

Silvio Peroni and David Shotton. The SPAR ontologies. In Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web – ISWC 2018*, page 119–136, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00668-6.

Grischka Petri. Der fall dürer vs. raimondi. vasaris erfindung. In Birgit Ulrike Münch, Andreas Tacke, Markwart Herzog, and Sylvia Heudecker, editors, *Fälschung – Plagiat – Kopie: Künstlerische Praktiken in der Vormoderne*, volume 1 of *Kunsthistorisches Forum Irsee*, pages 52–69. Michael Imhof Verlag, Petersberg, 2014.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language Models as Knowledge Bases?, September 2019. URL http://arxiv.org/abs/1909.01066. doi: 10.48550/arXiv.1909.01066.

James Phoenix and Mike Taylor. *Prompt Engineering for Generative AI*. "O'Reilly Media, Inc.", May 2024. ISBN 978-1-0981-5340-3.

Davide Picca. From intelligent to wise machines: Why a poem is worth more than 1 million tweets. *Informatik Spektrum*, 43(1):28–39, February 2020. ISSN 1432-122X. doi: 10.1007/s00287-020-01245-8.

Elena Pierazzo. *Digital Scholarly Editing: Theories, Models and Methods*. Ashgate, Farnham, 2015. ISBN 978-1-4724-1211-9.

Michael Piotrowski. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2012. doi: 10.2200/S00436ED1V01Y201207HLT017.

Michael Piotrowski. Accepting and modeling uncertainty. In Andreas Kuczera, Thorsten Wübbena, and Thomas Kollatz, editors, *Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten*, volume 4 of *Zeitschrift*

für digitale Geisteswissenschaften / Sonderbände. Herzog August Bibliothek Wolfenbüttel, Wolfenbüttel, 2019. doi: 10.17175/sb004_006a. URL https://zfdg.de/sb004_006.

Michael Piotrowski. Model, Corpus, Interpretation: Elements of Computational Hermeneutics. In Dennis Möbus, Christian Nawroth, Almut Leh, Uta Störl, and Matthias Hemmje, editors, *Digital Hermeneutics II: Sources, Analysis, Interpretation, Annotation, and Curation*, pages 35–50, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-08697-6. doi: 10.1007/978-3-032-08697-6_2.

Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. Agentic large language models, a survey. *Journal of Artificial Intelligence Research*, 84, 2025. doi: 10.1613/jair.1.18675. URL https://jair.org/index.php/jair/article/view/18675.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, 2022. doi: 10.18653/v1/2022.emnlp-main.731.

Fina Polat, Ilaria Tiddi, and Paul Groth. Testing prompt engineering methods for knowledge extraction from text. *Semantic Web*, 16(2):SW–243719, January 2025. ISSN 1570-0844. doi: 10.3233/SW-243719.

Polifonia Project. MEETUPS: A web interface for exploring historical musical encounters. https://polifonia.kmi.open.ac.uk/meetups/, 2023.

Lisa Pon. *Rewriting Vasari*, pages 261–275. Ashgate, 2014. ISBN 978-1-315-61301-7.

María Poveda-Villalón, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. OOPS! (OntOlogy Pitfall Scanner!): An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems*, 10(2):7–34, 2014. doi: 10.4018/ijswis.2014040102.

S. S. Pradhan, E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. OntoNotes: A unified relational semantic representation. In *International Conference on Semantic Computing (ICSC 2007)*, page 517–526, 2007. doi: 10.1109/ICSC.2007.83.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–40. Association for Computational Linguistics, 2012.

George A. Prenosil, Thilo K. Weitzel, Sandra C. Bello, Clemens Mingels, Giulia Manzini, Lorenz P. Meier, Kuang-Yu Shi, Axel Rominger, and Ali Afshar-Oromieh. Neuro-symbolic

AI for auditable cognitive information extraction from medical reports. *Communications Medicine*, 5(1):491, 2025. doi: 10.1038/s43856-025-01194-x. URL https://www.nature.com/articles/s43856-025-01194-x.

Valentina Presutti, Enrico Daga, Aldo Gangemi, and Eva Blomqvist. eXtreme design with content ontology design patterns. In *Proceedings of the 2009 International Conference on Ontology Patterns - Volume 516*, WOP'09, page 83–97, Aachen, DEU, 2009. CEUR-WS.org.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.494. URL https://aclanthology.org/2023.emnlp-main.494/.

Emily Pugh. Art History Now: Technology, Information, and Practice. *International Journal for Digital Art History*, 4:3.47–3.59, 2019. ISSN 2363-5401. doi: 10.11588/dah.2019.4.63448. URL https://ahnp.ub.uni-heidelberg.de/journals/dah/article/view/63448.

James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning*. O'Reilly Media, 2012.

M. Ross Quillian. Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12(5):410–430, 1967. doi: https://doi.org/10.1002/bs.3830120511. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bs.3830120511.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115. doi: 10.48550/arXiv.2412.15115.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL https://openai.com/index/language-unsupervised/. Preprint.

Stephen Ramsay. *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press, December 2011. ISBN 978-0-252-03641-5. doi: 10.5406/illinois/9780252036415.001.0001. URL https://doi.org/10.5406/illinois/9780252036415.001.0001.

Lance Ramshaw and Mitch Marcus. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*, 1995. URL https://aclanthology.org/W95-0107/.

Babak Ranjgar, Abolghasem Sadeghi-Niaraki, Maryam Shakeri, and Soo-Mi Choi. An ontological data model for points of interest (POI) in a cultural heritage site. *Heritage Science*, 10(1):13, 2022. doi: 10.1186/s40494-021-00635-9.

Heikki Rantala, Petri Leskinen, Lilli Peura, and Eero Hyvönen. Representing and searching associations in cultural heritage knowledge graphs using faceted search. In Angelo Salatino, Mehwish Alam, Femke Ongenae, Sahar Vahdati, Anna Lisa Gentile, Tassilo Pellegrini, and Shufan Jiang, editors, *Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI: Proceedings of the 20th International Conference on Semantic Systems (SEMANTiCS 2024)*, volume 60 of *Studies on the Semantic Web*, pages 420–435, Amsterdam, The Netherlands, 2024a. IOS Press. doi: 10.3233/SSW240033.

Heikki Rantala, Eljas Oksanen, Frida Ehrnsten, and Eero Hyvönen. Publishing numismatic public finds on the semantic web for digital humanities research – CoinSampo Linked Open Data service and semantic portal. In Oleksandra Bruns, Andrea Poltronieri, Lise Stork, and Tabea Tietz, editors, *Proceedings of the First International Workshop of Semantic Digital Humanities (SemDH 2024) co-located with the European Semantic Web Conference 2024 (ESWC 2024)*, volume 3724 of *CEUR Workshop Proceedings*, Hersonissos, Greece, 2024b. CEUR-WS.org. URL `https://ceur-ws.org/Vol-3724/paper3.pdf`.

Malte Rehbein. Digitalisierung. In Fotis Jannidis, Hubertus Kohle, and Malte Rehbein, editors, *Digital Humanities: Eine Einführung*, pages 179–198. J.B. Metzler, Stuttgart, 2017. ISBN 978-3-476-05446-3. doi: 10.1007/978-3-476-05446-3_12. URL `https://doi.org/10.1007/978-3-476-05446-3_12`.

Giulia Renda, Marco Grasso, and Marilena Daquino. From ontology design to user-centred interfaces for music heritage. In *Proceedings of the XII Convegno Annuale AIUCD (AIUCD 2023)*, 2023. doi: 10.5281/zenodo.11203264. URL `https://arxiv.org/abs/2306.12973`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning, 2016. URL `https://arxiv.org/abs/1606.05386`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: high-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Pat Riva, Patrick Le Bœuf, and Maja Žumer. IFLA library reference model: A conceptual model for bibliographic information. Technical report, International Federation of Library Associations and Institutions, 2017. URL `https://repository.ifla.org/handle/123456789/40`.

Rubin Maurice Rubin Patricia Lee. *Giorgio Vasari: Art and History*. Yale University Press, 1995. ISBN 978-0-300-04909-1. URL https://books.google.de/books?id=-X4mnRJaJ_UC.

Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, Upper Saddle River, NJ, 3 edition, 2010a. ISBN 978-0-13-604259-4.

Stuart J. Russell and Peter Norvig. What is AI? In *Artificial Intelligence: A Modern Approach*, chapter 1.1, pages 1–5. Prentice Hall, Upper Saddle River, NJ, 3 edition, 2010b. ISBN 978-0-13-604259-4.

Julia Röttgermann. The collection of eighteenth-century French novels 1751–1800. *Journal of Open Humanities Data*, 10:31, 2024. doi: 10.5334/johd.201.

Gernot Saalmann. The epistemological foundations of scientific knowledge. *Transcience*, 11 (2):106–122, 2020. URL https://www2.hu-berlin.de/transcience/Vol11_No2_106_122.pdf.

Patrick Sahle. What is a Scholarly Digital Edition? In Matthew James Driscoll and Elena Pierazzo, editors, *Digital Scholarly Editing: Theories and Practices*, pages 19–39. Open Book Publishers, Cambridge, 2016. doi: 10.11647/OBP.0095.01. URL https://books.openedition.org/obp/3382.

Sahar Salimpour, Lei Fu, Kajetan Rachwał, Pascal Bertrand, Kevin O'Sullivan, Robert Jakob, Farhad Keramat, Leonardo Militano, Giovanni Toffetti, Harry Edelman, and Jorge Peña Queralta. Towards Embodied Agentic AI: Review and Classification of LLM- and VLM-Driven Robot Autonomy and Interaction, 2025. URL https://arxiv.org/abs/2508.05294.

Robert Sanderson, Paolo Ciccarese, and Benjamin Young. Web annotation vocabulary, February 2017. URL https://www.w3.org/TR/annotation-vocab/.

C. Santini, M. A. Tan, O. Bruns, T. Tietz, E. Posthumus, and H. Sack. Knowledge Extraction for Art History: the Case of Vasari's The Lives of The Artists (1568). In *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022)*. CEUR-WS, 2022. URL https://u-pad.unimc.it/handle/11393/340813.

Cristian Santini, Mary Ann Tan, Oleksandra Bruns, Tabea Tietz, Etienne Posthumus, and Harald Sack. Guidelines for the annotation of ExtrART: Evaluation dataset for entity extraction from the lives of the artists (1550), May 2023. URL https://doi.org/10.5281/zenodo.7991340. doi: 10.5281/zenodo.7991340.

Cristian Santini, Nele Garay, Etienne Posthumus, and Harald Sack. The art of relations. In *Digital Humanities im deutschsprachigen Raum 2024 (DHd 2024)*, Passau, Germany, 2024.

Cristian Santini, Laura Melosi, and Emanuele Frontoni. Named Entity Recognition in Historical Italian: The Case of Giacomo Leopardi's Zibaldone, May 2025. URL http://arxiv.org/abs/2505.20113. doi: 10.48550/arXiv.2505.20113.

Cristian Santini, Marieke Van Erp, and Mehwish Alam. It's All About the Confidence: An Unsupervised Approach for Multilingual Historical Entity Linking using Large Language Models, January 2026. URL http://arxiv.org/abs/2601.08500. doi: 10.48550/arXiv.2601.08500.

Sunita Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1(3):261–377, 2008. doi: 10.1561/1900000003.

Advait Sarkar. Gluing pizza, eating rocks, and counting rs in strawberry: The discursive social function of stupid AI answers. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, CHIWORK '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713842. doi: 10.1145/3729176.3729189. URL https://doi.org/10.1145/3729176.3729189.

Bruno Sartini. IICONGRAPH: Improved Iconographic and Iconological Statements in Knowledge Graphs. In Albert Meroño Peñuela, Anastasia Dimou, Raphaël Troncy, Olaf Hartig, Maribel Acosta, Mehwish Alam, Heiko Paulheim, and Pasquale Lisena, editors, *The Semantic Web*, pages 57–74, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-60635-9. doi: 10.1007/978-3-031-60635-9_4.

Bruno Sartini, Sofia Baroncini, Marieke van Erp, Francesca Tomasi, and Aldo Gangemi. ICON: An ontology for comprehensive artistic interpretations. *J. Comput. Cult. Herit.*, 16(3), August 2023. ISSN 1556-4673. doi: 10.1145/3594724. URL https://doi.org/10.1145/3594724.

Andrea Schimmenti, Stefano De Giorgis, Fabio Vitali, and Marieke van Erp. Old Reviews, New Aspects: Aspect Based Sentiment Analysis and Entity Typing for Book Reviews with LLMs. In Mehwish Alam, Andon Tchechmedjiev, Jorge Gracia, Dagmar Gromann, Maria Pia di Buono, Johanna Monti, and Maxim Ionov, editors, *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 266–276, Naples, Italy, September 2025. Unior Press. ISBN 978-88-6719-333-2. URL https://aclanthology.org/2025.ldk-1.27/.

Andrea Schimmenti, Valentina Pasqual, Fabio Vitali, and Marieke van Erp. Knowledge graphs generation from cultural heritage texts: combining LLMs and ontological engineering for scholarly debates. *Journal of Documentation*, 82(7):206–250, March 2026. ISSN 0022-0418. doi: 10.1108/JD-07-2025-0203.

Marius Schlegel and Kai-Uwe Sattler. Capturing end-to-end provenance for machine learning pipelines. *Information Systems*, 132:102495, 2025. doi: 10.1016/j.is.2024.102495.

Eva Schlotheuber and Frank Bösch. Forum: Source Criticism in the Digital Age. ancillary historical skills as an essential competency for historical scholarship and associated disciplines. H-Soz-Kult, November 2015. URL https://www.hsozkult.de/debate/id/fddebate-132290. 15.11.2015.

Susan Schreibman. A companion to digital humanities, 2004. URL https://companions.digitalhumanities.org/DH/.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, et al. The Prompt Report: A Systematic Survey of Prompting Techniques, 2024. URL https://arxiv.org/abs/2406.06608. doi: 10.48550/arXiv.2406.06608.

Stefan Schweter and Johannes Baiter. Towards robust named entity recognition for historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 96–103. Association for Computational Linguistics, 2019.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, 2024. URL https://arxiv.org/abs/2310.11324.

Guido Seddone. LLMs and the increasing role of the humanities in the digital age. *AI & SOCIETY*, November 2025. ISSN 1435-5655. doi: 10.1007/s00146-025-02778-w. URL https://doi.org/10.1007/s00146-025-02778-w.

Isabel Segura-Bedmar et al. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, 2013.

Murray Shanahan. Talking about large language models. *Communications of the ACM*, 67(2): 68–79, 2024. doi: 10.1145/3624724.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06647-8. URL https://doi.org/10.1038/s41586-023-06647-8.

A. Shao. New sources of inaccuracy? a conceptual framework for studying AI hallucinations. *Harvard Kennedy School Misinformation Review*, 2025. doi: 10.37016/mr-2020-

157. URL https://misinforeview.hks.harvard.edu/article/new-sources-of-inaccuracy-a-conceptual-framework-for-studying-ai-hallucinations/.

Robert L J Shaw, Katalin Suba, Tomáš Hampejs, and David Zbíral. Syntactic-semantic capture of historical texts as a platform for source-critical analysis: telling the story of a premodern heresy trial with computer-assisted semantic text modelling (castemo). *Digital Scholarship in the Humanities*, 41(2):1001–1020, 06 2026. ISSN 2055-7671. doi: 10.1093/llc/fqag016. URL https://doi.org/10.1093/llc/fqag016.

Boheng Sheng, Jiacheng Yao, Meicong Zhang, and Guoxiu He. Dynamic Chunking and Selection for Reading Comprehension of Ultra-Long Context in Large Language Models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31857–31876, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1538. URL https://aclanthology.org/2025.acl-long.1538/.

Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. Generative multimodal entity linking. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7654–7665, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.676/.

Cogan Shimizu, Quinn Hirt, and Pascal Hitzler. MODL: A modular ontology design library, 2019. URL https://arxiv.org/abs/1904.05405.

Seiji Shimizu, Hisada Shohei, Yutaka Uno, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. Exploring LLM Annotation for Adaptation of Clinical Information Extraction Models under Data-sharing Restrictions. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14678–14694, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.757. URL https://aclanthology.org/2025.findings-acl.757/.

Darshana Shimpi and Sangita Chaudhari. An overview of graph databases. In *Proceedings of the International Conference in Recent Trends in Information Technology and Computer Science (ICRTITCS)*, volume ISSN 0975–8887, pages 16–22, 2012.

Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. A survey of data provenance in e-science. *ACM SIGMOD Record*, 34(3):31–36, 2005. doi: 10.1145/1084805.1084812.

Olaf Simons. FactGrid: A Database for Historians, 2024. URL https://database.factgrid.de/.

Amit Singhal. Introducing the knowledge graph: things, not strings, May 2012. URL https://blog.google/products/search/introducing-knowledge-graph-things-not/.

Ray Smith. An overview of the tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, 2007. doi: 10.1109/ICDAR.2007.4376991.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

C. M. Sperberg-McQueen and Lou Burnard, editors. *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Text Encoding Initiative, Chicago and Oxford, 1994.

Manu Sporny, Gregg Kellogg, and Markus Lanthaler. JSON-LD 1.0: A JSON-based serialization for linked data. W3c recommendation, W3C, January 2014. URL https://www.w3.org/TR/json-ld/.

Matteo Stefanini et al. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Image Analysis and Processing – ICIAP 2019*, pages 729–740, Cham, 2019. Springer International Publishing.

Matthias Steup and Ram Neta. Epistemology. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2025 edition, 2025.

Spencer Dean Stewart and Sanskriti Sinha. Retrieving information from unstructured historical sources using large language models. *Computational Humanities Research*, 1:e17, 2025. doi: 10.1017/chr.2025.10019. URL https://www.cambridge.org/core/product/845FDD146FF8A9283A32BFF0C3160F5F.

Sebastian Stober. Possibilities of source documentation and disclosure for generative AI systems. White paper, Otto-von-Guericke-University Magdeburg, Artificial Intelligence Lab, February 2025.

Ombretta Strafforello, Derya Soydaner, Michiel Willems, Anne-Sofie Maerten, and Stefanie De Winter. Have Large Vision-Language Models Mastered Art History?, September 2025. URL http://arxiv.org/abs/2409.03521. doi: 10.48550/arXiv.2409.03521.

Alessandra Stramiglio, Andrea Schimmenti, Valentina Pasqual, Marieke van Erp, Francesco Sovrano, and Fabio Vitali. Explicit vs. Implicit Biographies: Evaluating and Adapting LLM

Information Extraction on Wikidata-Derived Texts, September 2025. URL http://arxiv.org/abs/2509.14943. doi: 10.48550/arXiv.2509.14943.

Rudi Studer, V. Richard Benjamins, and Dieter Fensel. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1–2):161–197, 1998. doi: 10.1016/S0169-023X(97)00056-6.

Studies in Digital History and Hermeneutics. Studies in digital history and hermeneutics (book series). Book series, De Gruyter Oldenbourg. URL https://www.degruyterbrill.com/serial/sdhh-b/html. Series editors: Andreas Fickers, Valérie Schafer, Sean Takats and Gerben Zaagsma. ISSN 2629-4540.

Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. The NeOn methodology framework: A scenario-based methodology for ontology development. *Applied Ontology*, 10(2):107–145, 2015. ISSN 1570-5838. doi: 10.3233/AO-150145.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In Franck Dernoncourt, Daniel Preoțiuc-Pietro, and Anastasia Shimorina, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.91. URL https://aclanthology.org/2024.emnlp-industry.91/.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. *The Penn Treebank: An Overview*, page 5–22. Springer Netherlands, Dordrecht, 2003. ISBN 978-94-010-0201-1. doi: 10.1007/978-94-010-0201-1_1. URL https://doi.org/10.1007/978-94-010-0201-1_1.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. Association for Computational Linguistics, 2019.

Barbara Tillett. What is FRBR? a conceptual model for the bibliographic universe. *The Australian Library Journal*, 54(1):24–30, February 2005. ISSN 0004-9670. doi: 10.1080/00049670.2005.10721710.

Joshua Tint, Som Sagar, Aditya Taparia, Kelly Raines, Bimsara Pathiraja, Caleb Liu, and Ransalu Senanayake. ExpressivityBench: Can LLMs Communicate Implicitly?, February 2026. URL http://arxiv.org/abs/2411.08010. doi: 10.48550/arXiv.2411.08010.

Alba Morales Tirado, Jason Carvalho, Paul Mulholland, and Enrico Daga. Musical meetups: a knowledge graph approach for historical social network analysis. In Mehwish Alam, Cassia

Trojahn, Sven Hertling, Catia Pesquita, Christian Aebeloe, Hidir Aras, Amr Azzam, Juan Cano, John Domingue, Simon Gottschalk, Olaf Hartig, Katja Hose, Sabrina Kirrane, Pasquale Lisena, Francesco Osborne, Philipp Rohde, Luc Steels, Ruben Taelman, Aisling Third, Ilaria Tiddi, and Rima Türker, editors, *ESWC 2023 Workshops and Tutorials. Semantic Methods for Events and Stories (SEMMES)*, volume 3443. CEUR Workshop Proceedings (CEUR-WS.org), 2023. URL `https://oro.open.ac.uk/88720/`.

Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 1–4. Association for Computational Linguistics, 2002.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 142–147. Association for Computational Linguistics, 2003.

Francesca Tomasi. Digital editions between embedded markup and external representation. a case study: Vespasiano da bisticci's letters. In *Quaderni Digilab*, volume 2 of *Letterature, linguaggi e culture*, Università di Roma, 2012. La Sapienza. doi: 10.7357/DigiLab-24.

Francesca Tomasi. Vespasiano da bisticci, lettere, 2013. URL `http://vespasianodabisticciletters.unibo.it`.

Arthur Tompkins, editor. *Provenance Research Today: Principles, Practice, Problems*. Lund Humphries, London, 2020. ISBN 978-1-84822-276-2.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023. URL `https://arxiv.org/abs/2302.13971`. doi: 10.48550/arXiv.2302.13971.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yannis Tzitzikas, Michalis Mountantonakis, Pavlos Fafalios, and Yannis Marketakis. CIDOC-CRM and machine learning: A survey and future research. *Heritage*, 5(3):1612–1636, 2022. ISSN 2571-9408. doi: 10.3390/heritage5030084. URL `https://www.mdpi.com/2571-9408/5/3/84`.

Brian E. Ulicny, Jakub J. Moskal, Mieczyslaw M. Kokar, Keith Abe, and John Kei Smith. *Inference and Ontologies*, page 167–199. Springer International Publishing, Cham, 2014. ISBN 978-3-319-11391-3. doi: 10.1007/978-3-319-11391-3_9. URL https://doi.org/10.1007/978-3-319-11391-3_9.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72: 1385–1470, 2021. doi: 10.1613/jair.1.12752.

Miles J. Unger. *Michelangelo: A Life in Six Masterpieces*. Simon & Schuster, New York, 2014. ISBN 9781451678741.

Jacopo Urbani, Spyros Kotoulas, Eyal Oren, and Frank van Harmelen. Scalable distributed reasoning using MapReduce. In *Proceedings of the 8th International Semantic Web Conference (ISWC 2009)*, volume 5823 of *Lecture Notes in Computer Science*, pages 634–649. Springer, 2009. doi: 10.1007/978-3-642-04930-9_40.

Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. GERBIL: General entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 1133–1143, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741626. URL https://doi.org/10.1145/2736277.2741626.

Mahsa Vafaie. German national socialist injustice on the semantic web: from archival records to a knowledge graph. In K. Taylor and A. Zimmermann, editors, *Proceedings of the Doctoral Consortium at ISWC 2024 co-located with the 23rd International Semantic Web Conference (ISWC 2024)*, volume 3884 of *CEUR Workshop Proceedings*, Baltimore, MD, USA, November 2024. CEUR-WS. URL https://publikationen.bibliothek.kit.edu/1000179065.

Herbert Van de Sompel, Michael L. Nelson, Carl Lagoze, and Simeon Warner. Resource harvesting within the OAI-PMH framework. *D-Lib Magazine*, 10(12), 2004. doi: 10.1045/december2004-vandesompel.

Henri van de Waal. *Iconclass: an iconographic classification system*. North-Holland, Amsterdam, 1974–1985.

Giorgio Vasari. *Le vite de' più eccellenti pittori, scultori e architettori*. Giunti, Florence, Italy, 2 edition, 1568.

Giorgio Vasari. *Lives of the Most Eminent Painters, Sculptors, and Architects*. Henry G. Bohn, London, 1850/1852. URL https://digi.ub.uni-heidelberg.de/diglit/vasari1850ga.

Giorgio Vasari. *Das Leben des Pontormo*. Number 4 in Edition Giorgio Vasari. Verlag Klaus Wagenbach, Berlin, 1 edition, April 2004. ISBN 978-3-8031-5023-3. URL https://www.wagenbach.de/buecher/edition-giorgio-vasari.html.

Giorgio Vasari. *Das Leben des Tizian*. Number 8 in Edition Giorgio Vasari. Verlag Klaus Wagenbach, Berlin, 1 edition, March 2005. ISBN 978-3-8031-5027-1. URL https://www.wagenbach.de/buecher/edition-giorgio-vasari.html.

Giorgio Vasari. *Lives of the Most Eminent Painters, Sculptors and Architects*. Project Gutenberg, 2008. URL https://onlinebooks.library.upenn.edu/webbin/gutbook/lookup?num=25326.

Giorgio Vasari. *Das Leben des Michelangelo*. Number 26 in Edition Giorgio Vasari. Verlag Klaus Wagenbach, Berlin, 1 edition, September 2009. ISBN 978-3-8031-5045-5. URL https://www.wagenbach.de/buecher/edition-giorgio-vasari.html.

Giorgio Vasari. *Das Leben des Sandro Botticelli, Filippino Lippi, Cosimo Rosselli und Alesso Baldovinetti*. Number 27 in Edition Giorgio Vasari. Verlag Klaus Wagenbach, Berlin, 1 edition, February 2010. ISBN 978-3-8031-5046-2. URL https://www.wagenbach.de/buecher/edition-giorgio-vasari.html.

Giorgio Vasari. *Das Leben des Filippo Lippi, des Pesello und Francesco Peselli, des Andrea del Castagno und Domenico di Michelino*. Number 35 in Edition Giorgio Vasari. Verlag Klaus Wagenbach, Berlin, 1 edition, September 2011. ISBN 978-3-8031-5054-7. URL https://www.wagenbach.de/buecher/edition-giorgio-vasari.html.

Giorgio Vasari. *Das Leben des Verrocchio und der Gebrüder Pollaiuolo*. Number 36 in Edition Giorgio Vasari. Verlag Klaus Wagenbach, Berlin, 1 edition, May 2012a. ISBN 978-3-8031-5055-4. URL https://www.wagenbach.de/buecher/edition-giorgio-vasari.html.

Giorgio Vasari. *Das Leben des Brunelleschi und des Alberti*. Number 37 in Edition Giorgio Vasari. Verlag Klaus Wagenbach, Berlin, 1 edition, May 2012b. ISBN 978-3-8031-5056-1. URL https://www.wagenbach.de/buecher/edition-giorgio-vasari.html.

Giorgio Vasari. *Das Leben des Domenico Ghirlandaio und des Gherardo di Giovanni*. Number 42 in Edition Giorgio Vasari. Verlag Klaus Wagenbach, Berlin, 1 edition, April 2014. ISBN 978-3-8031-5061-5. URL https://www.wagenbach.de/buecher/edition-giorgio-vasari.html.

Giorgio Vasari. *Das Leben des Cimabue, des Giotto und des Pietro Cavallini*. Number 45 in Edition Giorgio Vasari. Verlag Klaus Wagenbach, Berlin, 1 edition, September 2015. ISBN 978-3-8031-5064-6. URL https://www.wagenbach.de/buecher/edition-giorgio-vasari.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Mark von Rosing, Wim Laurier, and Simon M. Polovina. *The Value of Ontology*, page 91–99. Morgan Kaufmann, Boston, January 2015. ISBN 978-0-12-799959-3. doi: 10.1016/B978-0-12-799959-3.00006-9. URL https://www.sciencedirect.com/science/article/pii/B9780127999593000069.

Denny Vrandečić. *Ontology Evaluation*, page 293–313. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-92673-3. doi: 10.1007/978-3-540-92673-3_13. URL https://doi.org/10.1007/978-3-540-92673-3_13.

Denny Vrandečić and Markus Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Jörg Waitelonis and Harald Sack. Towards exploratory video search using linked data. *Multimedia Tools and Applications*, 59(2):645–672, 2012. ISSN 1573-7721. doi: 10.1007/s11042-011-0733-1.

Nicholas Walker, Amalie Trewartha, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. The impact of domain-specific pre-training on named entity recognition tasks in materials science. *SSRN Electronic Journal*, 2021. doi: 10.2139/ssrn.3950755.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGPT a good NLG evaluator? a preliminary study, 2023a. URL https://arxiv.org/abs/2303.04048.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023b. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yue Wang and Man Zhang. CIDOC CRM-Based knowledge graph construction for cultural heritage using large language models. *Applied Sciences*, 15(22), 2025. ISSN 2076-3417. doi: 10.3390/app152212063. URL https://www.mdpi.com/2076-3417/15/22/12063.

Zhengxiang Wang, Jordan Kodner, and Owen Rambow. Exploring Limitations of LLM Capabilities with Multi-Problem Evaluation. In Aleksandr Drozd, João Sedoc, Shabnam Tafreshi, Arjun Akula, and Raphael Shu, editors, *The Sixth Workshop on Insights from Negative Results in NLP*, pages 121–140, Albuquerque, New Mexico, May 2025a. Association for Computational Linguistics. ISBN 979-8-89176-240-4. doi: 10.18653/v1/2025.insights-1.12. URL https://aclanthology.org/2025.insights-1.12/.

Zichong Wang, Zhibo Chu, Thang Viet Doan, Shiwen Ni, Min Yang, and Wenbin Zhang. History, development, and principles of large language models: an introductory survey. *AI and Ethics*, 5(3):1955–1971, June 2025b. ISSN 2730-5961. doi: 10.1007/s43681-024-00583-7. URL https://doi.org/10.1007/s43681-024-00583-7.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL http://arxiv.org/abs/2201.11903. doi: 10.48550/arXiv.2201.11903.

John C. Welchman. *Invisible Colors: A Visual History of Titles*. Yale University Press, New Haven, 1997. URL https://archive.org/details/invisiblecolorsv0000welc.

Toni Weller. Introduction: History in the digital age. In Toni Weller, editor, *History in the Digital Age*, pages 1–20. Routledge, London, 2013.

Jevin D. West and Carl T. Bergstrom. Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118(15):e1912444117, 2021. doi: 10.1073/pnas.1912444117.

Mitchell Whitelaw. Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly*, 9(1), 2015. URL http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html.

Roel J. Wieringa. *Research Goals and Research Questions*, page 13–23. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014a. ISBN 978-3-662-43839-8. doi: 10.1007/978-3-662-43839-8_2. URL https://doi.org/10.1007/978-3-662-43839-8_2.

Roel J. Wieringa. *Conceptual Frameworks*, page 73–91. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014b. ISBN 978-3-662-43839-8. doi: 10.1007/978-3-662-43839-8_8. URL https://doi.org/10.1007/978-3-662-43839-8_8.

Roel J. Wieringa. *Research Design*, page 121–133. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014c. ISBN 978-3-662-43839-8. doi: 10.1007/978-3-662-43839-8_11. URL https://doi.org/10.1007/978-3-662-43839-8_11.

Roel J. Wieringa. *Observational Case Studies*, page 225–245. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014d. ISBN 978-3-662-43839-8. doi: 10.1007/978-3-662-43839-8_17. URL https://doi.org/10.1007/978-3-662-43839-8_17.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), March 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18. URL https://doi.org/10.1038/sdata.2016.18.

Rhiannon Williams. Why Google's AI overviews gets things wrong. *MIT Technology Review*, May 2024. URL https://www.technologyreview.com/2024/05/31/1093019/why-are-googles-ai-overviews-results-so-bad/.

Florian Windhager, Saminu Salisu, Johannes Liem, and Eva Mayr. The knowledge graph as a data sculpture: Visualising arts and humanities data with maps, graphs, and sets over time. In Finn Dammann and Dominik Kremer, editors, *Geographical Research in the Digital Humanities*, pages 113–134. transcript Verlag, Bielefeld, 2024. ISBN 978-3-8394-6918-7. doi: 10.14361/9783839469187-007.

Robin Winkle, Manfred Stede, and Jörn Kreutel. Generative information extraction from biographical sources. In Diego Alves, Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva, Janis Pagel, and Stan Szpakowicz, editors, *Proceedings of the 10th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature 2026*, pages 311–322, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-373-9. doi: 10.18653/v1/2026.latechclfl-1.30. URL https://aclanthology.org/2026.latechclfl-1.30/.

Women Writers Project. Women writers online, 2023. URL https://www.wwp.northeastern.edu/.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing, pages 6397–6407. Association for Computational Linguistics, 2020.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A Survey of Uncertainty Estimation Methods on Large Language Models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1101. URL https://aclanthology.org/2025.findings-acl.1101/.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: a survey. *Frontiers of Computer Science*, 18(6):186357, November 2024. ISSN 2095-2236. doi: 10.1007/s11704-024-40555-y. URL https://doi.org/10.1007/s11704-024-40555-y.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.523.

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3091–3110, 2024. doi: 10.1109/TKDE.2024.3360454.

Yuzhao Yang, Genlang Chen, Binhua He, and Yan Zhao. Grounded knowledge graph extraction via LLMs: An anchor-constrained framework with provenance tracking. *Computers*, 15(3): 178, 2026. ISSN 2073-431X. doi: 10.3390/computers15030178.

Ruth Bernard Yeazell. *Picture Titles: How and Why Western Paintings Acquired Their Names*. Princeton University Press, 2015. ISBN 978-0-691-16527-1. doi: 10.2307/j.ctvc779dd. URL http://www.jstor.org/stable/j.ctvc779dd.

Robert K. Yin. *Case Study Research and Applications: Design and Methods*. SAGE Publications, Thousand Oaks, CA, 6 edition, 2018. ISBN 978-1-5063-3616-9.

J.D. Zamfirescu-Pereira, Raymond Y. Wong, Björn Hartmann, and Qian Yang. Why johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2023. doi: 10.1145/3544548.3581388.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.300.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. GLiNER2: An efficient multi-task information extraction system with schema-driven interface, 2025. URL https://arxiv.org/abs/2507.18546.

Marcia Lei Zeng. Smart data for digital humanities. *Journal of Data and Information Science*, 2 (1):1–12, 2017. doi: 10.1515/jdis-2017-0001.

Bohui Zhang, Filip Ilievski, and Pedro Szekely. Enriching wikidata with linked open data. In Lucie-Aimée Kaffee, Simon Razniewski, Gabriel Amaral, and Kholoud Saa Alghamdi, editors, *Proceedings of the Wikidata Workshop 2022 co-located with the 21st International Semantic Web Conference (ISWC 2022)*, volume 3262 of *CEUR Workshop Proceedings*, Hangzhou, China, October 2022. CEUR-WS.org. URL https://ceur-ws.org/Vol-3262/paper8.pdf.

Bohui Zhang, Valentina Anita Carriero, Katrin Schreiberhuber, Stefani Tsaneva, Lucía Sánchez González, Jongmo Kim, and Jacopo de Berardinis. OntoChat: a Framework for Conversational Ontology Engineering using Language Models, April 2024. URL http://arxiv.org/abs/2403.05921. doi: 10.48550/arXiv.2403.05921.

Qizheng Zhang, Michael Wornow, and Kunle Olukotun. Agentic plan caching: Test-time memory for fast and cost-efficient LLM agents. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 103270–103296. Curran Associates, Inc., 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9549f7d06700f0966d5f938f1d11022a-Paper-Conference.pdf.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu. Instruction Tuning for Large Language Models: A Survey. *ACM Comput. Surv.*, 58(7):169:1–169:36, January 2026. ISSN 0360-0300. doi: 10.1145/3777411. URL https://dl.acm.org/doi/10.1145/3777411.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models, 2025b. URL https://arxiv.org/abs/2506.05176. doi: 10.48550/arXiv.2506.05176.

Yutong Zhang, Dora Zhao, Jeffrey T. Hancock, Robert Kraut, and Diyi Yang. The rise of AI companions: How human-chatbot relationships influence well-being, 2025c. URL https://arxiv.org/abs/2506.12605.

Ze Yu Zhang, Arun Verma, Finale Doshi-Velez, and Bryan Kian Hsiang Low. Understanding the Relationship between Prompts and Response Uncertainty in Large Language Models, February 2025d. URL http://arxiv.org/abs/2407.14845. doi: 10.48550/arXiv.2407.14845.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. A Survey of Generative Information Extraction. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4840–4870, Abu Dhabi, UAE, January 2025e. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.324/.

Dong Zhao, Yadong Wang, Xiang Chen, Chenxi Wang, Hongliang Dai, Chuanxing Geng, Shengzhong Zhang, Shaoyuan Li, and Sheng-Jun Huang. Reflect then learn: Active prompting for information extraction guided by introspective confusion, 2025a. URL https://arxiv.org/abs/2508.10036.

Grace Zhao. Stanford misinformation expert admits to ChatGPT 'hallucinations' in court statement. *The Stanford Daily*, December 2024. URL https://stanforddaily.com/2024/12/04/hancock-admitted-to-ai-use/.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models, March 2025b. URL http://arxiv.org/abs/2303.18223. doi: 10.48550/arXiv.2303.18223.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Comput. Surv.*, 56(11), July 2024. ISSN 0360-0300. doi: 10.1145/3674501. URL https://doi.org/10.1145/3674501.

Hanwen Zheng, Sijia Wang, and Lifu Huang. A comprehensive survey on document-level information extraction. In Joel Tetreault, Thien Huu Nguyen, Hemank Lamba, and Amanda Hughes,

editors, *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 58–72, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.futured-1.6. URL https://aclanthology.org/2024.futured-1.6/.

Baohang Zhou, Zezhong Wang, Lingzhi Wang, Hongru Wang, Ying Zhang, Kehui Song, Xuhui Sui, and Kam-Fai Wong. DPDLLM: A Black-box Framework for Detecting Pre-training Data from Large Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 644–653, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.35. URL https://aclanthology.org/2024.findings-acl.35/.

Kang Zhou, Yuepei Li, Qing Wang, Qiao Qiao, and Qi Li. GenDecider: Integrating "none of the candidates" judgments in zero-shot entity linking re-ranking. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 239–245, Mexico City, Mexico, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.22. URL https://aclanthology.org/2024.naacl-short.22/.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. UniversalNER: Targeted distillation from large language models for open named entity recognition, 2024c. URL https://arxiv.org/abs/2308.03279.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimåmy Ba. Large language models are human-level prompt engineers, 2023. URL https://arxiv.org/abs/2211.01910.

Fangwei Zhu, Jifan Yu, Hailong Jin, Lei Hou, Juanzi Li, and Zhifang Sui. Learn to not link: Exploring NIL prediction in entity linking. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10846–10860, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.690. URL https://aclanthology.org/2023.findings-acl.690/.

Matthäus Zloch, Danilo Dessì, Jennifer D'Souza, Leyla Jael Castro, Benjamin Zapilko, Saurav Karmakar, Brigitte Mathiak, Markus Stocker, Wolfgang Otto, Sören Auer, and Stefan Dietze. Research knowledge graphs: the shifting paradigm of scholarly information representation, 2025. URL https://arxiv.org/abs/2506.07285.