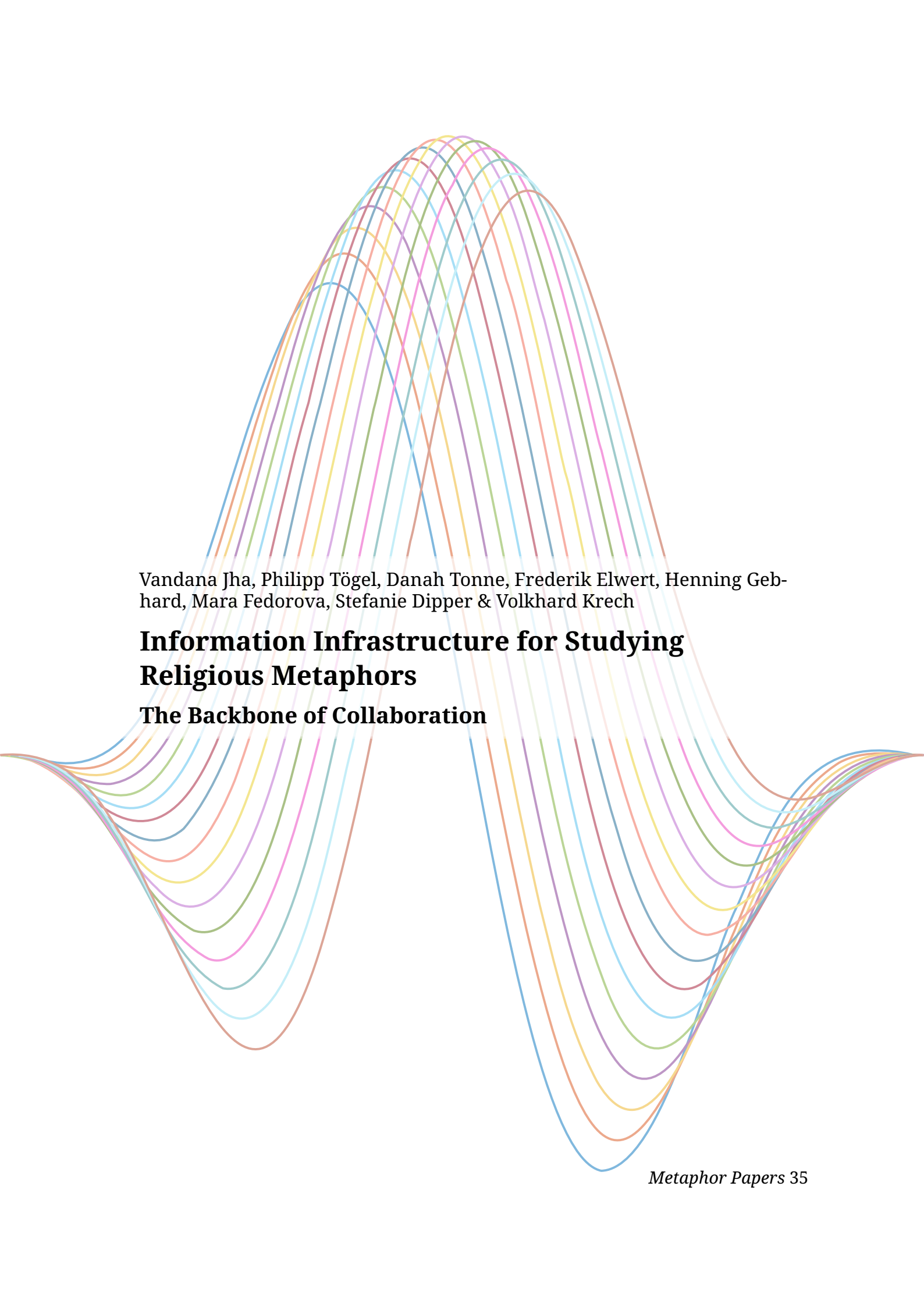




Vandana Jha, Philipp Tögel, Danah Tonne, Frederik Elwert, Henning Gebhard, Mara Fedorova, Stefanie Dipper & Volkhard Krech

**Information Infrastructure for Studying
Religious Metaphors
The Backbone of Collaboration**

Metaphor Papers is a Working Paper Series by the Collaborative Research Center 1475 “Metaphors of Religion”. In the *Metaphor Papers*, the CRC documents preliminary findings, work-in-progress and ongoing debates and makes them available for discussion.

Please cite as:

Vandana Jha, Philipp Tögel, Danah Tonne, Frederik Elwert, Henning Gebhard, Mara Fedorova, Stefanie Dipper & Volkhard Krech. “Information Infrastructure for Studying Religious Metaphors. The Backbone of Collaboration.” *Metaphor Papers* 35 (2026). <https://doi.org/10.46586/mp.485>.

© 2026 Vandana Jha, Philipp Tögel, Danah Tonne, Frederik Elwert, Henning Gebhard, Mara Fedorova, Stefanie Dipper, Volkhard Krech.

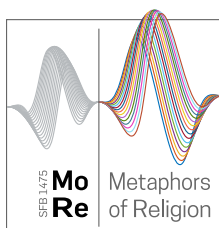
All *Metaphor Papers* are published under the Open Access CC-BY 4.0 International license: <https://creativecommons.org/licenses/by/4.0/>

ISSN 2942-0849

Ruhr-Universität Bochum / Karlsruher Institut für Technologie
Collaborative Research Center 1475 “Metaphors of Religion”

<https://sfb1475.ruhr-uni-bochum.de>
<https://omp.ub.rub.de/index.php/metaphorpapers>

The CRC “Metaphors of Religion” is funded by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) – SFB 1475 – Project ID 441126958.

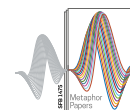


Funded by



Deutsche
Forschungsgemeinschaft

German Research Foundation



Information Infrastructure for Studying Religious Metaphors

The Backbone of Collaboration

Vandana Jha 

Karlsruhe Institute of Technology

Philipp Tögel 

Karlsruhe Institute of Technology

Danah Tonne 

Karlsruhe Institute of Technology

Frederik Elwert 

Ruhr University Bochum

Henning Gebhard 

Ruhr University Bochum

Mara Fedorova 

Ruhr University Bochum

Stefanie Dipper 

Ruhr University Bochum

Volkhard Krech 

Ruhr University Bochum

ABSTRACT Metaphors are present everywhere in both everyday language and in religious texts, playing a significant role in religious meaning-making processes. CRC 1475, “Metaphors of Religion,” is developing a shared infrastructure to facilitate metaphor analysis across diverse religious traditions and time periods. The infrastructure, based on the Web Annotation Data Model (WADM), stores and manages complex annotations resulting from metaphor analysis. The central infrastructure project (INF) within CRC 1475 provides the necessary frameworks and services to standardize metaphor analysis across 13 subprojects, which encompass a wide range of religious traditions. This paper introduces the shared infrastructure, highlighting key data modelling and infrastructure decisions, its development, and how it supports established metaphor-analytic approaches for metaphor identification and conceptual mapping. The infrastructure enables interdisciplinary collaboration, making the comparative study of religious metaphors possible on an unprecedented scale.

KEYWORDS Annotation, Infrastructure, Metaphor, Metaphor analysis, Religion

1. Introduction

Metaphors serve as tools that shape how we communicate abstract concepts and construct meaning in various domains, including religion, literature, and everyday communication. The use of metaphors is particularly significant in religious texts, which often

[1]

deal with complex and abstract notions such as divinity, morality, and transcendence. Religious metaphors serve as bridges between the known and the unknown, facilitating a deeper understanding of religious concepts (Krech, Karis, and Elwert 2023).

The CRC 1475 “Metaphors of Religion”¹ brings together 13 subprojects, each focusing on different religious traditions from around the world. These include Christianity, Islam, Judaism, Zoroastrianism, Jainism, Buddhism, and Daoism, with texts spanning from 3,000 BCE to the present day. The CRC works with texts written in multiple languages, ranging from ancient languages like Sumerian, Biblical Hebrew, and Sanskrit to modern languages such as German and English. The central question driving CRC 1475 is what role metaphorical language plays in religious meaning-making and how the metaphors used vary across different religions and cultures. Therefore, a comparative analysis across diverse religious texts is conducted, which faces a variety of challenges originating in the heterogeneity of the data (Artemov et al. 2024). [2]

To address these challenges, the CRC’s central infrastructure project (INF) is developing a shared data infrastructure that provides a unified framework for data representation, enrichment, and interoperability across all subprojects. The infrastructure is designed to support the continuous and collaborative annotation of heterogeneous textual sources and to enable comparative analyses across corpora, languages, and religious traditions. Central to this effort is the use of the Web Annotation Data Model (WADM, Young, Sanderson, and Ciccacese 2017), which offers a flexible and standards-based approach to represent complex stand-alone annotations and linking textual data with conceptual resources. By establishing common data models, reference frameworks, and interfaces, the CRC infrastructure accommodates diverse analytical perspectives while ensuring data consistency and reusability. This paper presents the design principles and architectural decisions underlying this infrastructure and demonstrates how it can accommodate established metaphor-analytic approaches, including conceptual metaphor theory (Lakoff and Johnson 2003) and Steen’s “Five Steps” (Steen 2011). [3]

2. CRC Requirements

While the diversity of subprojects and their subjects is essential to the research project, it creates substantial challenges for data integration, reuse, extensibility and comparison. The shared infrastructure must therefore satisfy a set of overarching requirements that emerge directly from the nature of the data and the collaborative research setting. These requirements concern textual heterogeneity, the need for systematic data enrichment, the establishment of a shared conceptual reference framework, and the possibility of cross-corpus analysis across languages, time-periods, and religious traditions. [4]

1 Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1475 – Project ID 441126958. CRC 1475 “Metaphors of Religion” consists of researchers from Ruhr University Bochum and Karlsruhe Institute of Technology.



Figure 1 Word Cloud Displaying Heterogeneity of Text Corpus

2.1 Heterogeneous Text Types and Sources

The primary requirement arises from the extreme heterogeneity of the textual data. [5] The CRC corpus spans multiple languages (as shown in figure 1), historical time-periods, and religious traditions, ranging from ancient religious texts to modern interpretations from *Reddit* forum data. These texts come in various formats, including PDF, Word, HTML, and CoNLL-U. Texts originate from different editorial traditions and scholarly workflows and exhibit variation in structure, completeness, orthography, and encoding practices. The encoding framework also has to handle special characters, quotations, gaps, and editorial interventions, which are frequent in historical and religious texts and require consistent treatment across corpora. Any shared infrastructure must therefore be capable of accommodating this diversity within a shared framework, but without privileging a single tradition or simplifying the data in ways that obscure relevant distinctions. In addition, the infrastructure needs to support fine-grained navigation and access to texts, not only at the document or chapter level, but also at the level of individual words. This requirement is essential for annotation, querying, and comparison across corpora. All texts are required to be normalized to ensure consistent handling of special characters, diacritics, and non-Latin scripts across corpora.

2.2 Enriching Text with Interpretation

A second core requirement concerns the progressive enrichment of textual data. Research in CRC 1475 is not limited to the analysis of raw texts; instead, texts are continuously enriched with metaphor identifications and conceptual mappings. This enrichment process is iterative and collaborative and unfolds over long periods of time. [6]

The infrastructure must therefore support the attachment of multiple layers of information to the same textual material without destabilizing the underlying texts. It must allow enrichment to grow incrementally, support multiple analyses of the same or overlapping spans of text, and ensure that newly added information remains traceable and reusable. Since enrichment in CRC 1475 is primarily realized through annotations, the infrastructure requires stable and fine-grained textual identifiers. Rather than relying on chapter indices or page numbers, all texts are therefore tokenized and assigned persistent word-level (or character-level in the case of logographic scripts such as Chinese) identifiers. These stable references ensure that minor textual corrections (e.g. typograph- [7]

ical errors) do not invalidate existing annotations and allow annotation layers to remain reliably anchored over time. This requirement implies systematic tokenization of all corpora as part of the data preparation process prior to annotation.²

2.3 Common Reference Framework

Comparative research across subprojects requires more than shared texts and annotations; it requires shared points of conceptual reference. Scholars working on different languages and religious traditions often use different terminologies to describe related phenomena. Without a common reference framework, annotations remain isolated within subprojects and cannot be meaningfully compared. [8]

Therefore, the infrastructure must support a controlled and extensible system of concepts that allows researchers to relate textual phenomena to shared abstract categories. Such a framework must be flexible enough to accommodate domain-specific distinctions while still enabling semantic alignment across corpora, languages, and analytical traditions. [9]

2.4 Cross-corpus Analysis

Finally, the CRC requires explicit support for cross-corpus analysis. The central research questions concern similarities, differences, and patterns of metaphor usage across religious traditions, historical periods, and linguistic contexts. Addressing these questions requires the ability to aggregate, compare, and explore data across corpora that were analyzed and annotated by independent teams of researchers using the same basic methodology. [10]

The infrastructure must therefore enable analysis that goes beyond individual texts or subprojects. It must support filtering, aggregation, and comparison across multiple dimensions, such as language, concept, religious tradition, or historical period, and present these relationships in forms that are interpretable by scholars from different disciplinary backgrounds. [11]

2.5 Modular Infrastructure and API-Based Integration

Beyond textual heterogeneity and enrichment requirements, the infrastructure must support a modular architecture in which individual components can operate independently while remaining interoperable. The breadth of CRC 1475 requirements—text storage, annotation management, vocabulary curation, and cross-corpus analysis—cannot be adequately addressed by a single monolithic system. Each task imposes distinct technical and conceptual demands. [12]

Rather than building a tightly coupled system, CRC 1475 therefore adopts a service-oriented design in which repository, annotation, vocabulary, and analysis components [13]

2 For implementation details and examples see following sections, esp. 4.1.

communicate via standardized HTTP-based REST interfaces. These APIs provide controlled, machine-readable access to texts, annotations, and conceptual resources, enabling flexible integration of internally developed tools and external services. This modular approach ensures maintainability, extensibility, and long-term sustainability, while allowing individual components to evolve independently without disrupting the overall infrastructure.

3. Design Principles

The requirements outlined above inform a set of design principles that guide the development of the CRC 1475 infrastructure. All design decisions are guided by the FAIR data principles.³ Texts, annotations, and vocabularies are made findable through structured metadata and persistent identifiers, accessible via standardized HTTP-based APIs, interoperable through the use of community standards such as TEI-XML, RDF (Resource Description Framework) and SKOS (Simple Knowledge Organization System), and reusable through open formats, open-source software components, clear licensing, and modular design. These principles ensure that data produced within the CRC remains usable both within and beyond the project's lifespan. [14]

3.1 Common Representation of Heterogeneous Texts

Given the diversity of texts and scholarly practices within the CRC, maintaining separate technical serializations and encoding practices for each subproject would severely limit interoperability and comparative analysis. To address these challenges, all texts are converted into a common TEI-XML (Burnard and Bauman 2007) format. TEI-XML is chosen because it provides a well-established, flexible, and community-driven standard for representing textual data in the digital humanities. It supports rich structural and semantic markup while allowing project-specific customisations through extensions. This balance between standardization and customization is essential for supporting collaborative, cross-corpus research (Jha et al. 2025). Tokenization is applied at the word level, except for logographic scripts (e.g. Chinese), where character-level tokenization ensures accurate linguistic representation. Text normalization uses NFC to standardize special characters across languages, including accented letters and non-Latin scripts. [15]

3.2 Data Enrichment via Annotations

Stand-off annotations were chosen over inline markup to support continuous enrichment, collaborative workflows, and reuse. By decoupling annotations from the textual data, the infrastructure allows new analytical layers to be added without altering the underlying corpus and enables the annotation of overlapping XML elements. This design facilitates versioning, querying, and aggregation across corpora and tools. [16]

Annotations are modelled using the WADM, which provides a standardized frame- [17]

3 <https://www.go-fair.org/fair-principles/>.

work for representing annotations as first-class web resources. WADM defines explicit relationships between annotation bodies, targets, and metadata, ensuring that metaphor identifications and conceptual mappings remain machine-readable, interoperable, and persistently addressable.

3.3 RDF as a Unifying Data Model

To enable integration across textual data, annotations, and conceptual vocabularies, the infrastructure adopts RDF as a unifying data model. RDF enables the explicit modelling of relationships between texts, annotations, and concepts while remaining flexible enough to accommodate evolving research questions. [18]

By representing enrichment information as Linked Open Data, RDF supports interoperability, extensibility, and semantic clarity across the CRC (Tögel et al. 2025). Crucially, RDF allows analytical information to be expressed independently of the textual representation itself, enabling aggregation, cross-corpus analysis, and integration with external semantic resources. [19]

3.4 Concepts and Thesaurus

Shared vocabularies and thesauri are essential in a large, interdisciplinary CRC. Without a common conceptual framework, analysis results would remain isolated within individual subprojects, limiting interoperability and cross-corpus comparison (Ide and Pustejovsky 2017; Heath and Bizer 2011). In the CRC infrastructure, concepts are modelled using SKOS (Miles and Bechhofer 2009), an RDF-based vocabulary standard, facilitating semantic consistency and enabling meaningful comparison across languages, traditions, and analytical perspectives. [20]

A conceptual thesaurus provides a shared reference framework for standardizing terms and conceptual categories across diverse subprojects. This facilitates consistent interpretation of abstract and domain-specific phenomena and supports cross-lingual and cross-project comparability. The thesaurus is organized hierarchically and based on existing data from the Semantic Domains project.⁴ It is continually extended by incorporating domain-specific concepts relevant to CRC 1475.⁵ For example, an Arabic metaphor using the word *ṣirāṭ* (road) and a Chinese metaphor using *dao* (path) can both be linked to the same thesaurus concept “Way, route” with the URI <https://w3id.org/MoRe-SFB1475/CT/concepts/3792706109>. [21]

To support scholars in identifying suitable concepts for linking their analytical results, a semantic search interface was developed. This interface combines traditional hierarchical browsing via SKOSMOS⁶ (Suominen et al. 2015) with a Bidirectional Encoder Representations from Transformers (BERT)-based semantic similarity model. While SKOSMOS provides structured access to the controlled vocabulary through its SKOS [22]

⁴ <https://semdom.org/>

⁵ The work-in-progress conceptual thesaurus is available at <https://w3id.org/MoRe-SFB1475/CT/concepts>.

⁶ <https://skosmos.org/>

endpoints, the BERT model enables similarity-based retrieval of concept suggestions beyond exact lexical matches. By integrating vector-based semantic search with controlled vocabulary browsing, the system facilitates efficient concept discovery and promotes consistent conceptual alignment across heterogeneous datasets.

4. Architecture

The architecture of CRC 1475 infrastructure operationalizes the requirements and design principles outlined above through a set of modular and interoperable components as shown by the central part, “Metaphor Workbench” in figure 2. Each component addresses a clearly defined subset of requirements—textual heterogeneity, continuous enrichment, shared conceptual reference, and cross-corpus metaphor analysis—while relying on standardized data models and interfaces to ensure interoperability across the infrastructure. To keep integration effort manageable, the infrastructure relies on common programming languages and widely adopted technologies, avoiding unnecessary heterogeneity at the implementation level. This design choice supports maintainability and lowers the barrier for future extensions or reuse of individual components in other research contexts. [23]

4.1 Text Repository

A centralized text repository is a foundational component of CRC 1475 infrastructure, as it provides a single, reliable point of access to all textual data produced and curated across the 13 subprojects. Given the diversity of sources, languages, and editorial practices, such a repository is essential for ensuring consistency, long-term preservation, and interoperability within the infrastructure. [24]

From an infrastructural perspective, the repository must satisfy several core requirements. First, it must support TEI-encoded texts as the primary content format, ensuring compatibility with established digital humanities standards and enabling rich structural and semantic markup. Second, it must manage structured metadata, covering both administrative metadata (e.g. provenance, licensing, versioning) and content-related metadata (e.g. language, period, subproject affiliation). Treating metadata and content as structured, machine-readable entities is crucial for discoverability, reuse, and automated processing. Third, the repository must provide HTTP-based access to all stored resources via standardized, well-documented APIs. Such interfaces are a prerequisite for a modular infrastructure, allowing annotation tools, analysis services, and visualization components to interact with the data without tight coupling. [25]

Beyond functional requirements, several design principles guided the choice of repository solution. The infrastructure prioritizes the use of standards and community recommendations for both data and metadata, ensuring long-term sustainability and interoperability beyond the CRC. Furthermore, all components are based on open-source software, combining newly developed modules with established solutions. This approach not only [26]

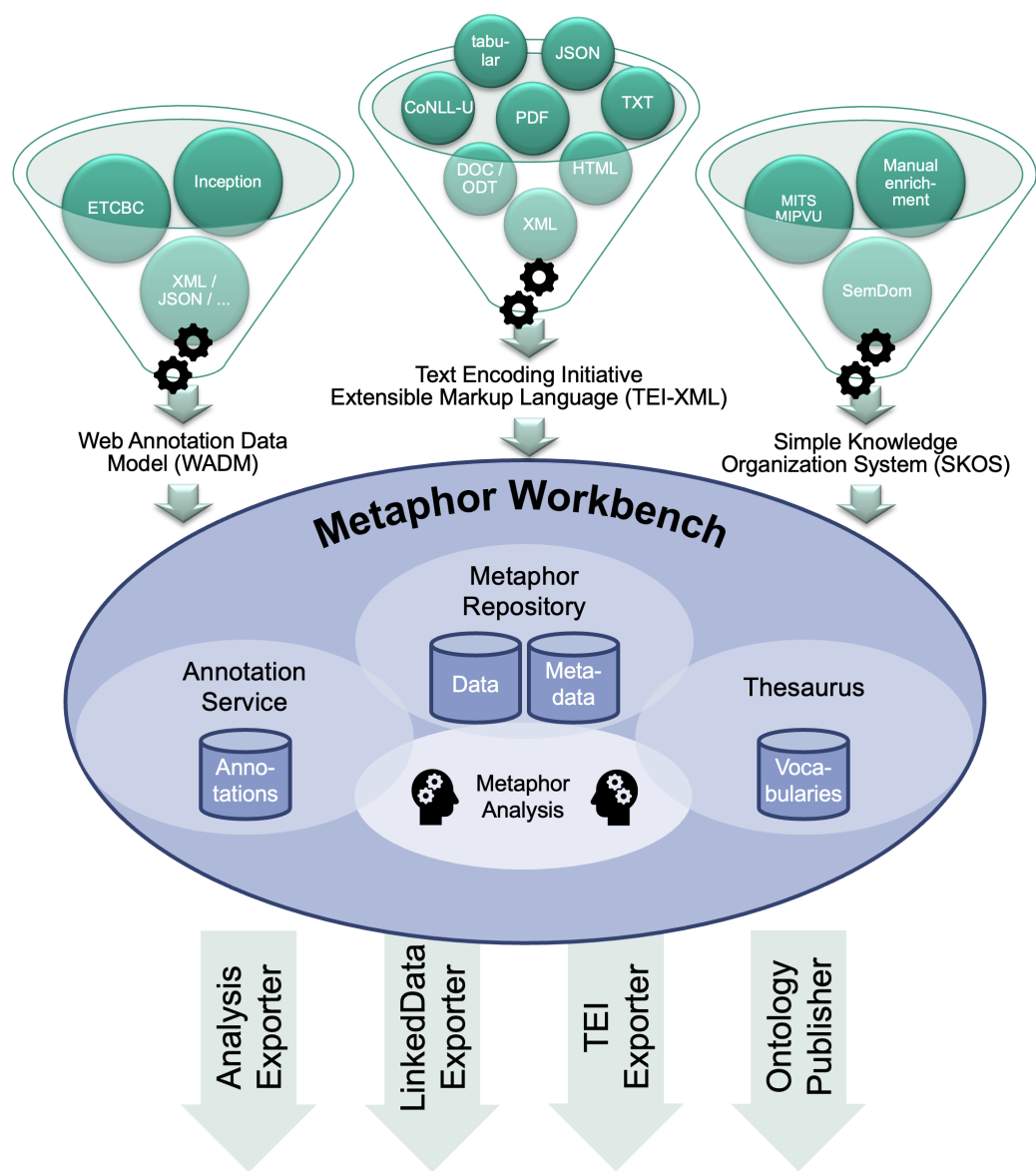


Figure 2 Architecture diagram of CRC 1475 infrastructure framework

promotes transparency and reusability but also allows the infrastructure to benefit from existing ecosystems and best practices.

As an implementation of these principles, CRC 1475 employs an instance of the KIT Data Manager Base Repo⁷ (Jejkal et al. 2014). The Base Repo stores texts and metadata as structured Digital Objects, assigns persistent identifiers, and exposes all resources via standardized HTTP-based interfaces, which serve as the primary access layers for downstream tools that actively work with the texts. Metadata and text content are maintained as separate but linked Digital Objects, allowing independent updates while preserving their conceptual unity. Stable identifiers ensure that minor textual corrections or updates do not invalidate existing references or annotations. In close collaboration with the subprojects, the INF team supports data ingestion, format conversion and metadata creation, ensuring that all texts enter the repository in a consistent and FAIR-compliant manner. [27]

Apart from the encoding of every token (word- or character-level depending on the language) to allow for the creation of stable targets, the following phenomena were especially relevant for the encoding of the texts (some are included in listing 1): [28]

- The segmentation of each document into chapters, sections etc. was encoded to enable navigation to a specific segment. [29]
- Quotes and references to literature or other source material were already marked in the data source in a machine- or human-readable way and got converted to TEI-XML.
- Due to the historical nature of some texts, incomplete or unreadable passages as well as editorial interventions had to be encoded using the EpiDoc guidelines (Elliot et al. 2017–2026).
- To encode different pronunciations of individual words, ruby-annotations were used.
- For some languages different representations of complete documents are available (i.e. sandhied/unsandhied Sanskrit texts and vocalized/unvocalized Hebrew texts) and have been encoded using a combination of various elements including choice-elements.⁸

```
<div n="De coniunctione" xml:id="c.7" type="chapter">
  <head><w xml:id="w.3035">De</w> <w xml:id="w.3036">coniunctione</w></head>
  <ab>
    <w xml:id="w.3037">Sequitur</w> <w xml:id="w.3038">de</w>
    <w xml:id="w.3039">coniunctione</w><pc xml:id="pc.604">.</pc>
    <w xml:id="w.3040">Debemus</w> <w xml:id="w.3041">enim</w>
    <w xml:id="w.3042">esse</w> <w xml:id="w.3043">deo</w>
    <w xml:id="w.3044">coniuncti</w><pc xml:id="pc.605">,</pc>
    <w xml:id="w.3045">ne</w> <w xml:id="w.3046">ab</w>
```

7 <https://github.com/kitt-data-manager/base-repo>

8 For a more detailed overview on the encoding see Jha et al. (2025).

```

<w xml:id="w.3047">eo</w> <w xml:id="w.3048">per</w>
<w xml:id="w.3049">peccatum</w> <w xml:id="w.3050">simus</w>
<w xml:id="w.3051">separati</w><pc xml:id="pc.606">.</pc>
<w xml:id="w.3052">Cui</w> <w xml:id="w.3053">fuimus</w>
<w xml:id="w.3054">in</w> <w xml:id="w.3055">baptismo</w>
<w xml:id="w.3056">coadunati</w> <w xml:id="w.3057">ita</w>
<pc xml:id="pc.607">,</pc> <w xml:id="w.3058">quod</w>
<w xml:id="w.3059">cum</w> <w xml:id="w.3060">eo</w>
<w xml:id="w.3061">simus</w> <w xml:id="w.3062">unum</w>
<w xml:id="w.3063">corpus</w><pc xml:id="pc.608">,</pc>
<w xml:id="w.3064">et</w> <w xml:id="w.3065">ipse</w>
<w xml:id="w.3066">est</w> <w xml:id="w.3067">caput</w>
<w xml:id="w.3068">nostrum</w> <w xml:id="w.3069">et</w>
<w xml:id="w.3070">nos</w> <w xml:id="w.3071">sumus</w>
<w xml:id="w.3072">membra</w> <w xml:id="w.3073">eius</w>
<pc xml:id="pc.609">.</pc> <quote><w xml:id="w.3074">Tollens</w>
<w xml:id="w.3075">ergo</w> <w xml:id="w.3076">membra</w>
<w xml:id="w.3077">Christi</w> <w xml:id="w.3078">faciam</w>
<w xml:id="w.3079">membra</w> <w xml:id="w.3080">meretricis</w>
<pc xml:id="pc.610">?</pc> <w xml:id="w.3081">Absit</w></quote>
<supplied resp="#steinbach.johannes">1. Kor 6,15</supplied>
<pc xml:id="pc.611">.</pc>
[...]
```

</ab>

</div>

Listing 1 Tokenized and TEI encoded version of “Donatus Mysticus” transcribed from the manuscript Augsburg, Universitätsbibliothek Augsburg, Codex II.1.4°20, f.88r-93r. by Johannes Steinbach.

4.2 Annotation Editor and Storage

The second architectural layer addresses the requirement for systematic, long-term enrichment of textual data. Rather than embedding analytical information directly within the textual representation, the infrastructure adopts a stand-off annotation approach, allowing multiple, independent layers of interpretation to coexist without altering the base texts. This separation is essential for supporting parallel analyses, iterative refinement, and continuous enrichment over time, while avoiding the well-known problem of overlapping markup structures that arises in inline XML-based annotation. Annotations serve as research data in their own right, capturing metaphor identifications and conceptual associations while remaining explicitly linked to stable textual targets. By decoupling annotations from the source texts, the infrastructure [30]

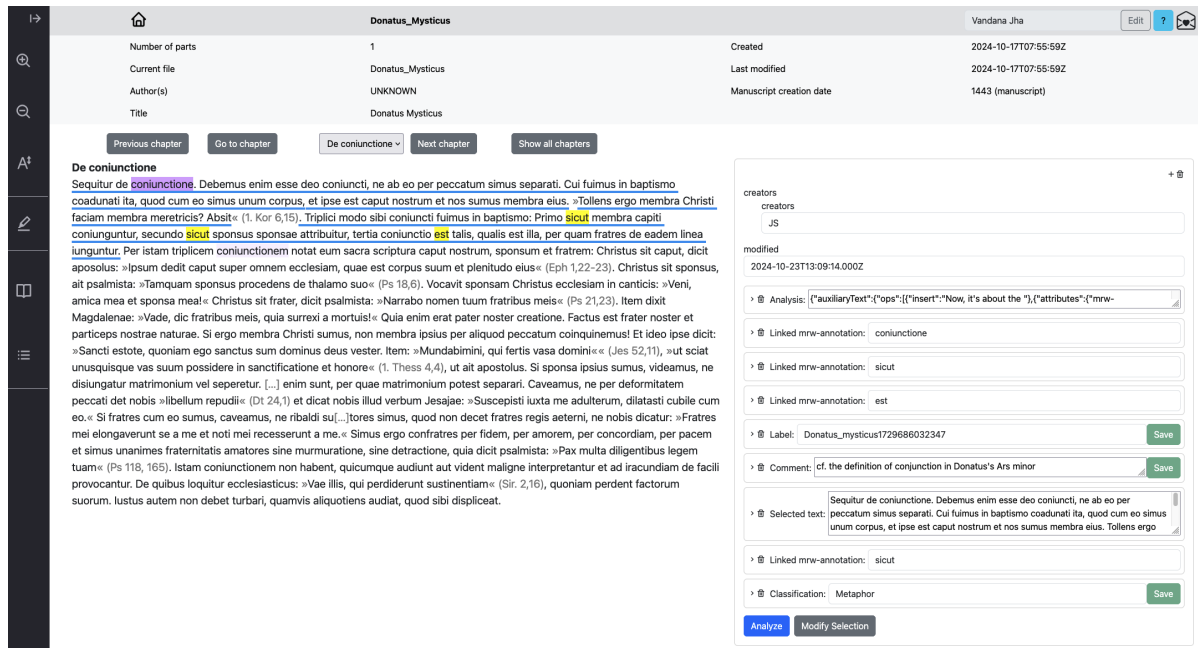


Figure 3 Text annotator of tAKITA illustrating flexible annotation capabilities

preserves textual integrity and enables reuse of both texts and annotations across analytical contexts.

This conceptual approach is realized through the combination of the tAKITA annotation editor (see figure 3)⁹ and a WAP-compliant annotation server¹⁰ (Tonne et al. 2019). tAKITA provides a graphical interface for creating, editing, and managing annotations, allowing researchers to reference fine-grained textual targets using stable token identifiers. The tAKITA plugin, Metrow (Metaphor Row, as illustrated in figure 4) serves as a tool for the complex analysis of individual metaphors. All annotations are transmitted and accessed via a REST interface fully compliant with the W3C Web Annotation Protocol (Sanderson 2017). Annotations are stored using the WADM (Young, Sanderson, and Ciccacese 2017), a W3C Recommendation and successor of the Open Annotation Data Model (Sanderson, Ciccacese, and Van de Sompel 2013). In this model, annotations are represented as a set of connected resources, explicitly defining relationships between targets (in our case textual segments) and bodies (annotation content). The use of RDF enables expressive modelling of complex annotation structures and supports flexible querying via SPARQL.

CRC 1475 leverages the expressiveness and flexibility offered by the WADM by including the results of the metaphor identification and analysis as a Body, which holds a SpecificResource containing a dctypes:Dataset. To specify why the Body is included in the annotation, values taken from the TaDiRAH taxonomy (<https://vocabs.dariah.eu/tadirah>, Borek et al. 2021) are reused (e.g. tadirah:analyzing in the case of the metaphor analysis Web Annotations; see listing 2). As all texts have been tokenized on a word level, the `xml:ids` of each word are used as "anchors" for an XPathSelector,

9 <https://github.com/kit-data-manager/takita>

10 <https://github.com/kit-data-manager/wap-server>

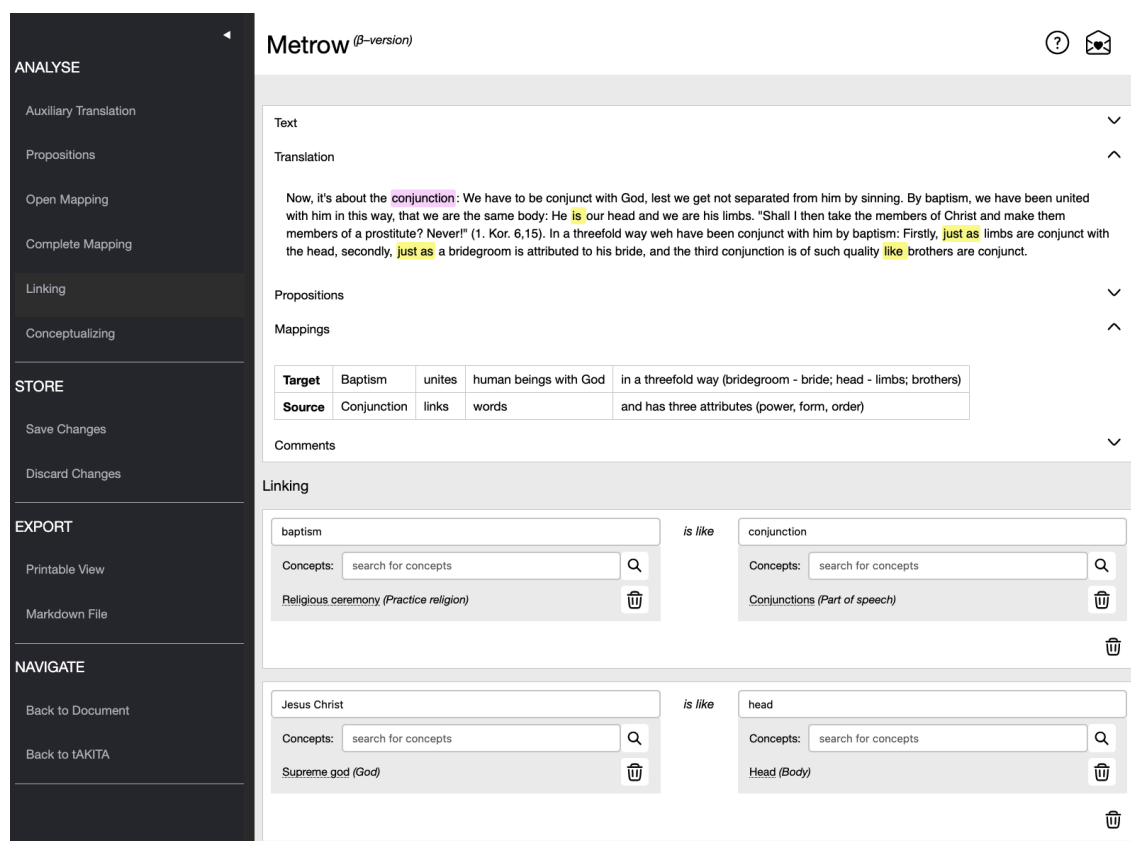


Figure 4 Metrow facilitating metaphor analysis

which is included in the annotations Target. An additional TextQuoteSelector allows for the verbatim storage of the selected text in the triple store (see listing 3) and the inclusion of the selected text in query results. Furthermore, it creates a fallback option for consumers unable to resolve the XPathSelector. Nevertheless, the XPathSelector is the primary way to retrieve the text, as it is more robust than the TextQuoteSelector if minor changes in the document are necessary [e.g. corrections of spelling mistakes; Tögel et al. (2025) describes our annotation model in detail].

This standards-based approach ensures interoperability, facilitates annotation exchange, and enables the integration and conversion of heterogeneous annotation sources—including annotations produced in external tools such as INCEPTION¹¹ (Klie et al. 2018) as well as pre-existing linguistic layers (e.g. part-of-speech tags, lemma or other corpus annotations)—into a unified Web Annotation representation. [33]

```
{ "body": [
  {
    "type": "SpecificResource",
    "purpose": "tadirah:analyzing",
    "source": {
      "id": "moredata:6fabb347123",
      "type": "Dataset",

```

11 <https://inception-project.github.io/>

```

"value": {
  "id": "moredata:6fabb347123MetaphorAnalysis",
  "type": "more:MetaphorAnalysis",
  "more:hasMapping": [
    {
      "more:hasSource": {
        "type": "more:DomainInstance",
        "id": "moredata:dbebd22e27dc1",
        "more:hasTerm": "conjunction",
        "more:hasConcept": [
          "https://w3id.org/MoRe-SFB1475/CT/concepts/1033334983"
        ]
      },
      "more:hasTarget": {
        "type": "more:DomainInstance",
        "id": "moredata:dbebd22e27dc2",
        "more:hasTerm": "baptism",
        "more:hasConcept": [
          "https://w3id.org/MoRe-SFB1475/CT/concepts/1998715950"
        ]
      }
    }
  ],
  "more:hasMRW": ["moredata:6fabb347123MRW"]
}
}
}]

```

Listing 2 Simplified embedded MetaphorAnalysis resource

```

{"target": [
  {
    "selector": {
      "type": "XPathSelector",
      "value": "id(\"w.3039\")"
    },
    "source": "https://example.org/texts/Donatus_mysticus.xml"
  },
  {
    "selector": {
      "type": "TextQuoteSelector",
      "exact": "coniunctione",

```

```

    "prefix": "Sequitur de ",
    "suffix": ". Debemus enim esse deo coniuncti"
  },
  "source": "https://example.org/texts/Donatus_mysticus.xml"
}
]]

```

Listing 3 Text selection

4.3 Vocabulary Editor and Publisher

To enable semantic alignment across subprojects, the infrastructure includes a dedicated component for managing shared conceptual resources. This layer provides a controlled vocabulary that functions as a common reference framework for annotation, ensuring that interpretations applied to texts from different languages, traditions, and periods remain semantically comparable (see [Dipper and Elwert 2024](#) for an in-depth presentation of our methodology and how the thesaurus is used). [34]

Entries in the thesaurus are represented as instances of `skos:Concept` within an RDF model, enabling hierarchical organization, multilingual labels, and explicit semantic relations between entries (see listing 4). This representation supports the consistent linking of annotations to abstract concepts rather than project-specific terminology, which is essential for cross-corpus and cross-lingual analysis. [35]

The thesaurus is collaboratively developed and maintained using VocBench¹² ([Stellato et al. 2020](#)), which provides a web-based graphical interface for editing and curating SKOS vocabularies. Editorial workflows support continuous extension and refinement of the conceptual framework during the research process. Curated vocabularies are then published via a SKOSMOS instance,¹³ making them accessible through standardized interfaces for both human users and machine-based tools (see figure 5). To keep the concepts findable even if the infrastructure moves, we are using w3ids. [36]

By separating vocabulary creation from vocabulary publication, this component ensures editorial flexibility while providing a stable and reusable semantic reference layer for annotation and analysis across the CRC infrastructure. [37]

¹² <http://vocbench.uniroma2.it/>

¹³ The work-in-progress conceptual thesaurus is available at <https://w3id.org/MoRe-SFB1475/CT/concepts> and the Metaphor Identification Tagset at <https://w3id.org/MoRe-SFB1475/MITS>.

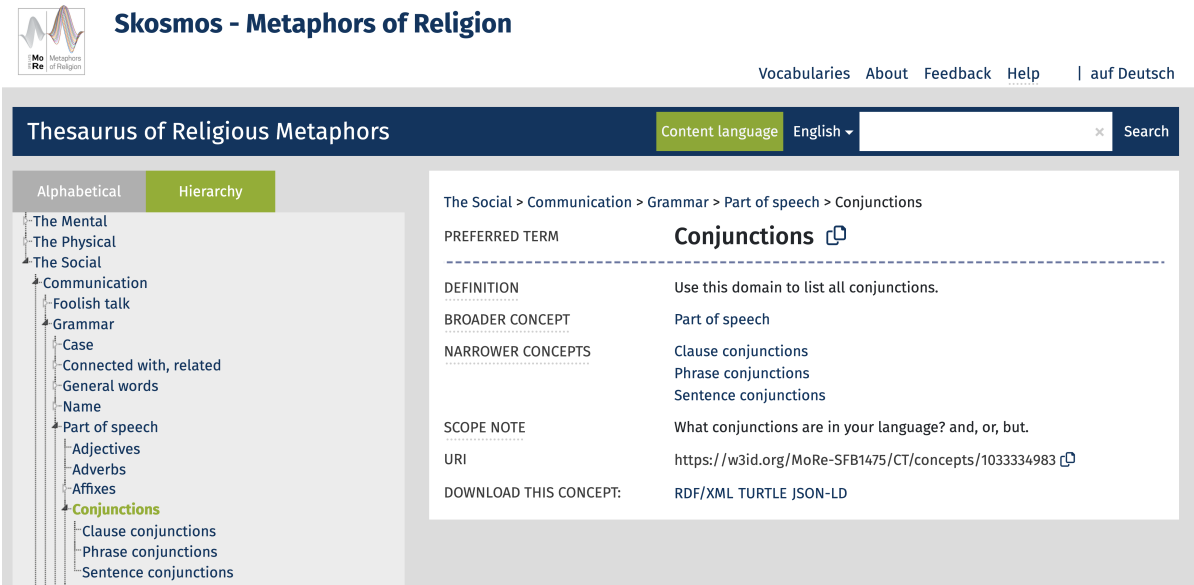


Figure 5 Conceptual Thesaurus published via SKOSMOS

```

@prefix : <https://w3id.org/MoRe-SFB1475/CT/concepts/> .
@prefix cc: <http://creativecommons.org/ns#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

```

```

<https://w3id.org/MoRe-SFB1475/CT/concepts> a skos:ConceptScheme ;
    dct:creator <https://www.wikidata.org/wiki/Q114613694> ;
    dct:description """Conceptual Thesaurus is a SKOS vocabulary/ontology for
annotating religious metaphors. It adapts and extends material from the
Semantic Domains (SemDom) project, originally developed as a tool for
dictionary creation by SIL International. This ontology is a derivative
work of SemDom and includes substantial modifications (re-organization,
additions, removals, and new properties). Content adapted from SemDom is
attributed to SIL International; all edits and new material are by the
CRC 1475 "Metaphors of Religion: Religious Meaning-Making in Language Use"
project.
"""@en ;
    dct:license <http://creativecommons.org/licenses/by-sa/4.0/> ;
    skos:altLabel "Semdom"@en ;
    skos:hasTopConcept :1355632831,
        :232319197,
        :293386450 ;
    skos:prefLabel "Semantic Domains"@en .

:1033334983 a skos:Concept ;
    skos:broader :3671287479 ;
    skos:definition "Use this domain to list all conjunctions."@en ;
    skos:inScheme <https://w3id.org/MoRe-SFB1475/CT/concepts> ;
    skos:notation "9.2.5"^^<https://w3id.org/MoRe-SFB1475/CT/TCPATH> ;
    skos:prefLabel "Conjunctions"@en ;
    skos:scopeNote "What conjunctions are in your language? and, or, but."@en .

:1000027404 a skos:Concept ;
    skos:broader :1033334983 ;
    skos:definition """Use this domain to list all phrase level
conjunctions--conjunctions that join two words within a phrase.
"""@en ;
    skos:inScheme <https://w3id.org/MoRe-SFB1475/CT/concepts> ;

```

```

skos:notation "9.2.5.1"^^<https://w3id.org/MoRe-SFB1475/CT/TCPath> ;
skos:prefLabel "Phrase conjunctions"@en ;
skos:scopeNote ""What conjunctions join two words within a phrase?
and, both...and, or, either...or, neither...nor, but.
""@en .

```

Listing 4 Excerpt of Conceptual Thesaurus in Turtle

4.4 Annotation Aggregator and Analyzer

The final architectural layer shown in figure 6 addresses the requirement for cross-corpus analysis and comparative exploration. Building on standardized textual representations, RDF-based annotations, and shared conceptual references, this component aggregates annotation data across corpora and subprojects. The Annotation Aggregator and Analyzer enables researchers to explore relationships between concepts, texts, and annotations across multiple dimensions, such as language, time-period and religious tradition. By operating on aggregated RDF data rather than individual text files, this component supports analyses that are not feasible within isolated subproject contexts. [38]

The infrastructure implements these requirements through the MED-Kit (Metaphor Exploration and Discovery Kit, see figure 6). MED-Kit visualizes networks of concepts, where nodes represent conceptual categories and edges represent annotated metaphorical relationships derived from aggregated RDF data (see figure 7). Scholars can filter and compare data across languages, subprojects, and text attributes. MED-Kit operates on aggregated Web Annotations, including textual fragments stored within annotation bodies, to provide a user-friendly interface for exploring relationships, conducting comparative analyses, and identifying patterns across the CRC corpus, directly addressing the project's central research questions. This layer demonstrates how the modular architecture enables higher-level analytical capabilities without imposing uniform analytical assumptions on individual subprojects. [39]

4.5 Export and Interoperability Layer

In addition to internal storage, enrichment, and analysis, the CRC 1475 infrastructure provides a set of export interfaces that support reuse and interoperability beyond the immediate project context. These exporters (shown in the lower part of figure 2) do not constitute independent analytical components; instead, they provide structured export mechanisms that transform the internally maintained RDF-based representations into standardized formats suitable for reuse in different scholarly tools (like MED-Kit), platforms, and research workflows. [40]

Depending on the research context, enriched data can be exported for quantitative and exploratory analysis, for integration into Linked Data environments, for reuse in TEI-based philological workflows, or for ontology-driven applications. By offering multiple export paths, the infrastructure avoids imposing a single mode of data consumption or [41]

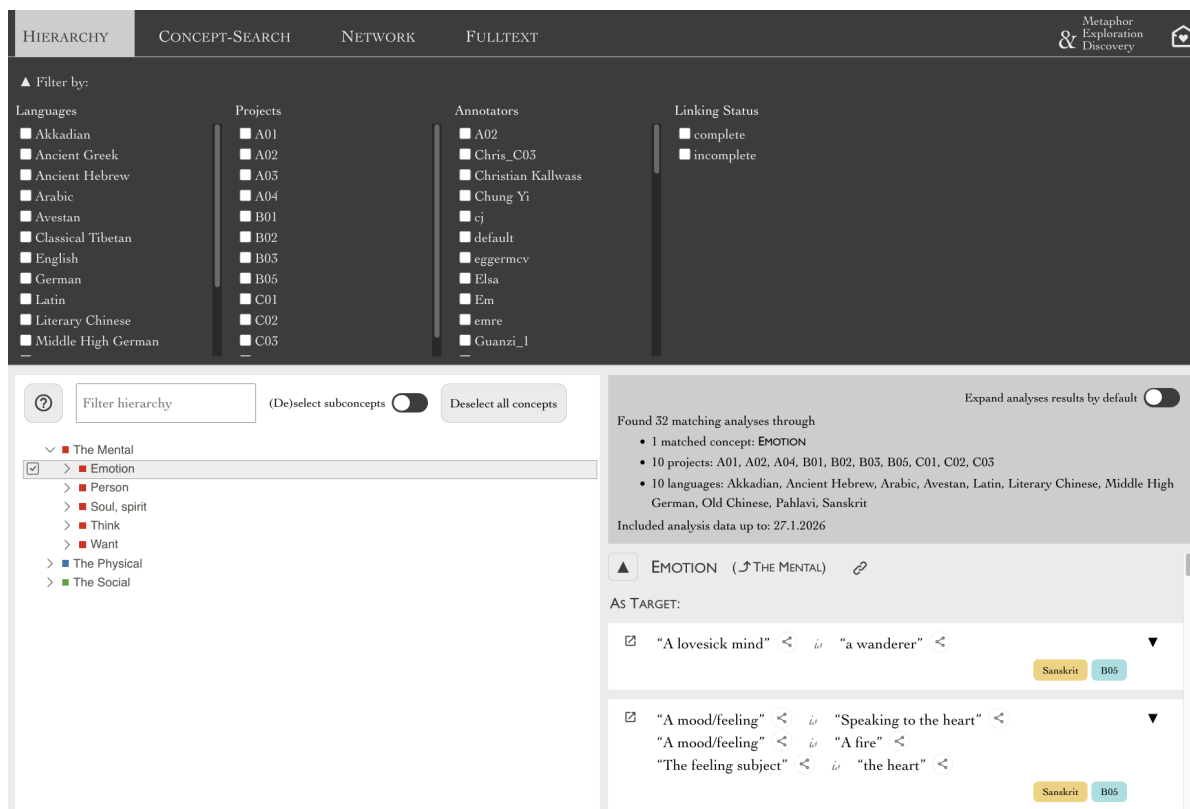


Figure 6 MED-Kit supporting annotation aggregation

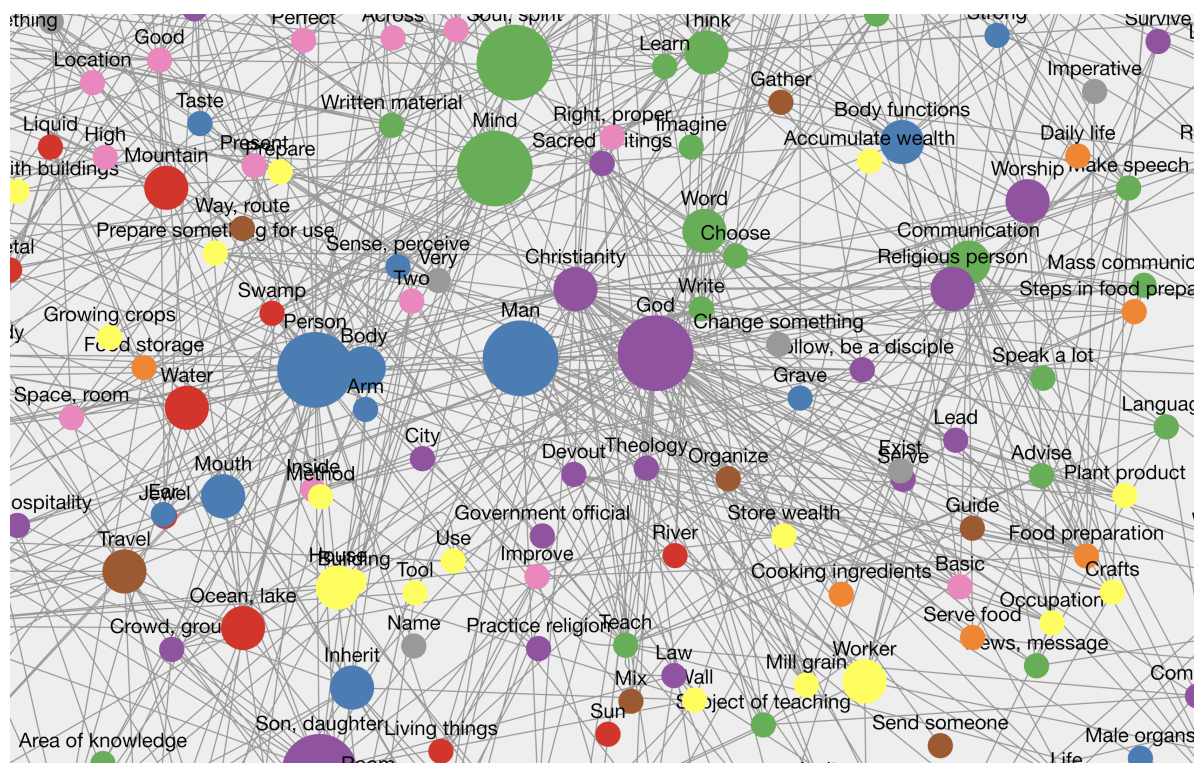


Figure 7 MED-Kit concept network visualization showing metaphorical mappings across corpora

analysis, while preserving consistent internal data models. This design directly supports the FAIR principles, in particular interoperability and reusability, and ensures that data produced within the CRC remains accessible and meaningful across disciplinary, technical, and institutional boundaries.

5. Results and Conclusion

The implementation of the CRC 1475 infrastructure demonstrates the feasibility of harmonizing heterogeneous religious corpora within a unified semantic framework and integrated texts from 13 subprojects into a common TEI-XML format, supporting annotation and cross-corpus exploration. Table 1 summarizes the dataset and annotation statistics, retrieved on 12.01.2026. These figures reflect the state of the corpus at the time of writing this paper and will continue to evolve as annotation progresses.

By using the WADM, it provides a flexible and powerful framework for conducting comparative analyses of metaphorical language across diverse religious traditions. The shared infrastructure and standardized methodology enable researchers to explore how metaphors function in meaning-making processes across time, languages, and cultures. The generated data and the open-access nature of the tools further enhance its value, promoting reusability and interoperability for future research. As the tools continue to be refined and expanded, they will help in advancing our understanding of metaphor in religious language, providing insights into the cognitive and cultural processes that shape human thought.

Table 1 Corpus and annotation statistics, retrieved on June 30, 2026.

Metric	Count
Texts	2,042
Tokens	81,378,026
Annotations	16,009
Concepts	1,823

CRC 1475 infrastructure demonstrates how heterogeneous textual corpora can be harmonized, enriched, and analysed using modular RDF-based components. By providing standardized data models, annotation frameworks, and visualization tools, the infrastructure supports cross-lingual and cross-project research, enabling advanced comparative analyses. The design decisions—modular architecture, WADM-based annotations, and a shared conceptual thesaurus—ensure flexibility, scalability, and reusability for future research.

References

- Artemov, Nikita, Elsa Kueppers, Sebastian Reimann, Lina Rodenhausen, and Alexandra Wiemann. 2024. "Taking MIPVU Further Around the World—and Through the Ages: MIPVU for Religion-Related Texts: Challenges and Benefits." *Metaphor Papers* 9. <https://doi.org/10.46586/mp.299>.
- Borek, Luise, Canan Hastik, Vera Khramova, Klaus Illmayer, and Jonathan D. Geiger. 2021. "Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR." In *Information Between Data and Knowledge*, 321–32. Schriften Zur Informationswissenschaft 74. Glückstadt: Werner Hülsbusch. <https://doi.org/10.5283/epub.44951>.
- Burnard, Lou, and Syd Bauman, eds. 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.10.2, September 2025. Charlottesville: TEI Consortium. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.
- Dipper, Stefanie, and Frederik Elwert. 2024. "Annotating Metaphorical Mappings. An Implementation of Steen's Five Step Method." *Metaphor Papers* 11. <https://doi.org/10.46586/mp.315>.
- Elliot, Tom, Gabriel Bodard, Elli Mylonas, Simona Stoyanova, Charlotte Tupman, and Scott Vanderbilt et. al. 2017–2026. *EpiDoc Guidelines: Ancient Documents in TEI XML (Version 9)*. <https://epidoc.stoa.org/gl/latest/>.
- Heath, Tom, and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space*. San Rafael: Morgan & Claypool.
- Ide, Nancy, and James Pustejovsky. 2017. *Handbook of Linguistic Annotation*. Dordrecht: Springer.
- Jejkal, Thomas, Alexander Vondrous, Andreas Kopmann, Rainer Stotzka, and Volker Hartmann. 2014. "KIT Data Manager: The Repository Architecture Enabling Cross-Disciplinary Research." In *Large-Scale Data Management and Analysis (LSDMA) - Big Data in Science*, edited by Christopher Jung and Achim Streit, 9–11. Karlsruhe: Karlsruhe Institute of Technology (KIT). <https://publikationen.bibliothek.kit.edu/1000043392>.
- Jha, Vandana, Philipp Tögel, Danah Tonne, Frederik Elwert, Henning Gebhard, Makar Fedorov, Stefanie Dipper, and Henny-Krahmer. 2025. "Text Corpus in Collaboration: A Balance Between Customized and Standardized Approach." In *FORGE 2025 — Forschungsdaten in Den Geisteswissenschaften: Daten Neu Denken. Konferenzabstracts*, 101. <https://doi.org/10.5281/zenodo.17178219>.
- Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation." In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. <https://aclanthology.org/C18-2002>.
- Krech, Volkhard, Tim Karis, and Frederik Elwert. 2023. "Metaphors of Religion. A Conceptual Framework." *Metaphor Papers* 1. <https://doi.org/10.46586/mp.282>.

- Lakoff, George, and Mark Johnson. 2003. *Metaphors We Live by*. 2nd ed. Chicago: University of Chicago Press.
- Miles, Alistair, and Sean Bechhofer. 2009. "SKOS Simple Knowledge Organization System Reference." <https://www.w3.org/TR/skos-reference/>.
- Sanderson, Robert. 2017. "Web Annotation Protocol." <https://www.w3.org/TR/annotation-protocol/>.
- Sanderson, Robert, Paolo Ciccarese, and Herbert Van de Sompel. 2013. "Designing the W3C Open Annotation Data Model." In *Proceedings of the 5th Annual ACM Web Science Conference*, 366–75. <https://doi.org/10.1145/2464464.2464474>.
- Steen, Gerard. 2011. "From Three Dimensions to Five Steps: The Value of Deliberate Metaphor." *Metaphorik.de* 21: 83–110. <https://doi.org/10.17185/dupublico/83794>.
- Stellato, Armando, Manuel Fiorelli, Andrea Turbati, Tiziano Lorenzetti, Willem Van Gemert, Denis Dechandon, Christine Laaboudi-Spoiden, et al. 2020. "VocBench 3: A Collaborative Semantic Web Editor for Ontologies, Thesauri and Lexicons." *Semantic Web* 11 (5): 855–81. <https://doi.org/10.3233/SW-200370>.
- Suominen, Osmo, Henri Ylikotila, Sini Pessala, Mikko Lappalainen, Matias Frosterus, Jouni Tuominen, Thomas Baker, Caterina Caracciolo, and Armin Retterath. 2015. "Publishing SKOS Vocabularies with Skosmos."
- Tögel, Philipp, Henning Gebhard, Frederik Elwert, Stefanie Dipper, Makar Fedorov, Vandana Jha, Volkhard Krech, and Danah Tonne. 2025. "Data-Rich Web Annotations: Embedding Datasets to Link Complex Metaphor Analyses with Their Textual Basis." In *Proceedings of the Second International Workshop of Semantic Digital Humanities (SemDH 2025) Co-Located with ESWC 2025*. CEUR Workshop Proceedings 4009. <https://doi.org/10.5445/IR/1000183935>.
- Tonne, Danah, Germaine Götzelmann, Philipp Hegel, Michael Krewet, Julia Hübner, Sibylle Söring, Andreas Löffler, Michael Hitzker, Markus Höfler, and Timo Schmidt. 2019. "Ein Web Annotation Protocol Server Zur Untersuchung Vormoderner Wissensbestände." In *DHd 2019 Digital Humanities: Multimedial & Multimodal. Konferenzabstracts*, 283–85. <https://doi.org/10.5281/zenodo.2596094>.
- Young, Benjamin, Robert Sanderson, and Paolo Ciccarese. 2017. "Web Annotation Data Model." <https://www.w3.org/TR/annotation-model/>.