Knowledge Graph enabled Curation and Exploration of Nuremberg's City Heritage

Tabea Tietz[1,2], Oleksandra Bruns[1,2], Sandra Göller[1], Matthias Razum[1], Danilo Dessì[1,2], and Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
firstname.lastname@fiz-karlsruhe.de
[2] Karlsruhe Institute of Technology, Institute AIFB, Germany

Abstract. An important part in European cultural identity relies on European cities and in particular on their histories and cultural heritage. Nuremberg, the home of important artists such as Albrecht Dürer and Hans Sachs developed into the epitome of German and European culture already during the Middle Ages. Throughout history, the city experienced a number of transformations, especially with its almost complete destruction during World War 2. This position paper presents TRANSRAZ, a project with the goal to recreate Nuremberg by means of an interactive 3D tool to explore the city's architecture and culture ranging from the 17th to the 21st century. The goal of this position paper is to discuss the ongoing work of connecting heterogeneous historical data from various sources previously hidden in archives to the 3D model using knowledge graphs for a scientifically accurate interactive exploration on the Web.

Keywords: Knowledge Graphs · History · Cultural Heritage .

1 Introduction

Preparing cultural heritage collections for exploration by a wide range of users with multidisciplinary backgrounds requires to integrate heterogeneous historical data into modern information systems to structure and curation. Knowledge graphs (KGs) have proven to be a reliable way of structuring data for exploration purposes. They enable to connect data within historical collections by means of ontologies, enrich them with external data sources like Wikidata or national authority files, and identify entities within collections unambiguously by using URIs as persistent identifiers. However, historical collections are challenging to integrate into a KG as the data quality varies and rarely one-fits-all-solutions can be applied.

Goal of this position paper is to introduce the efforts of the ongoing research project TRANSRAZ[3] in which heterogeneous historical data collections are

[3] https://www.fiz-karlsruhe.de/en/forschung/transraz

connected to an architectural 3D virtual research environment (VRE) using KGs to enable the exploration of the historic city of Nuremberg in different time periods ranging from the Middle Ages to the 21st century. Exploring city architectures along with the people living in it, their progress in technology, their craftmanships as well as arts and culture is highly relevant for many domains related to (digital) humanities. Nuremberg was one of the great European metropolises in the Middle Ages and beyond. It was the birthplace of renaissance artist Albrecht Dürer, who worked there all his life. The city developed into the epitome of German and European history and culture. Then, during the Second World War, the city was largely destroyed and only few buildings could be reconstructed. However, without a systematic and scientific reconstruction of the city in different time periods, this important part of the European cultural heritage will be forgotten.

This reconstruction of Nuremberg was first initiated with the TOPORAZ research project[7] in which a VRE was created that links a scholarly sound 3D model of the main market of the city of Nuremberg to a database in four different time layers. The project TRANSRAZ as presented in this paper builds on these efforts and extends the VRE from a city square to the entire historical city center with around 3.000 houses. Furthermore, a knowledge graph will be created to connect historical data acquired from archives to the exploration environment. KGs play a fundamental role in the curation of historical data sources and their integration with the 3D environment for this project. They allow to build a meaningful data model using open standards for exploration on the Web, and enable to connect to further related resources provided by e.g., Wikidata as well as galleries, libraries, archives and museums (GLAM). In this paper, the vision of using a KG for curating data connected to Nuremberg's history and making it available for research and education purposes as part of a 3D exploration tool will be explained including the challenges faced along the way.

The remainder of the paper is structured as follows. Section 2 introduces related work, in Section 3 the TRANSRAZ project as main use case is presented. Section 4 presents an overview of the data, the envisioned workflow and curation challenges, followed by a conclusion in Section 5.

2 Related Work

The interest in the representation and curation of the social, cultural, and geographical evolution of the past through digital systems has recently taken hold in several initiatives. Especially the exploration of urban spaces and their changes throughout time have been matter of interdisciplinary research projects for a few years. One of the most prominent and largest efforts in this domain was the *Time Machine*[4] project. It aimed to rebuild the history of European cities, not only by digitizing the archives, but also by opening the doors to modern technologies to shape the elements that represent their evolution. As an example, spatial and temporal historical information about places, people, and events of the

[4] `https://www.timemachine.eu/`

city of Amsterdam have been digitized and represented through the *Amsterdam Time Machine*[5]. This platform bridges data ranging from humanities to cultural heritage, and provides linked open data (LOD) to explore. An important feature provided by the *Amsterdam Time Machine* is the interlinking between the LOD and 3D models, which provides the possibility to explore the city of Amsterdam in space and time, and supports a better human understanding of its evolution. The work by [6] as part of the Urban Complexity Lab is highly relevant in the exploration of city scapes from the perspective of digital humanities, science communication and smart cities. Their visualizations are focused on the challenges and questions arising from social, cultural, and technological transformations.

The Timetraveler Berlin application for smartphones lets users experience historical multimedia content of the Berlin Wall by means of augmented reality. The application guides users on a GPS-based tour to historically significant locations in Berlin and displays stories of events that happened in history[10].

In comparison to previous attempts, TRANSRAZ will focus on automated systems to transform the data from its unstructured form to meaningful representations, and will provide data curation opportunities through ontologies and KGs. The project will carry out research to study methodologies to integrate and make sense of data coming from heterogeneous sources, and will provide formal representations of contents for advanced explorations that are not supported by current systems. Remarkable and timely examples of existing formal representations are ArCO [2], a KG which provides ontology patterns to link people, events, and places about Italian artifacts and document collections, Linked Stage Graph [9] which organizes and interconnects data about the Stuttgart State Theaters, and ArDO [11] an ontology to represent the dynamics of annotations of general archival resources. To achieve the set goals, one important part of research within TRANSRAZ focuses on the extraction of textual content from digitized images [1], the use of natural language processing tools to recognize various entities (e.g., people, places, and events) and relations among them, as well as the development, adaption, and mapping of ontologies to model the extracted information into KGs.

3 TRANSRAZ: Exploring Nuremberg's History in 3D

The use case of the presented approach is an exploration tool for the city of Nuremberg, provided in the ongoing project TRANSRAZ by means of an interactive 3D city model. Its goal is to provide scientifically accurate means of exploration for Nuremberg in various points of time ranging from the 17th to the 21st century. The research of urban spaces and art history usually starts from location-bound objects, e.g. buildings or streets. The 3D reconstruction and research of buildings, creating complexes and entire topographical spaces that no longer exist or have undergone major changes is now an established procedure in the digital humanities. The presented approach will provide a direct

[5] `https://amsterdamtimemachine.nl/`

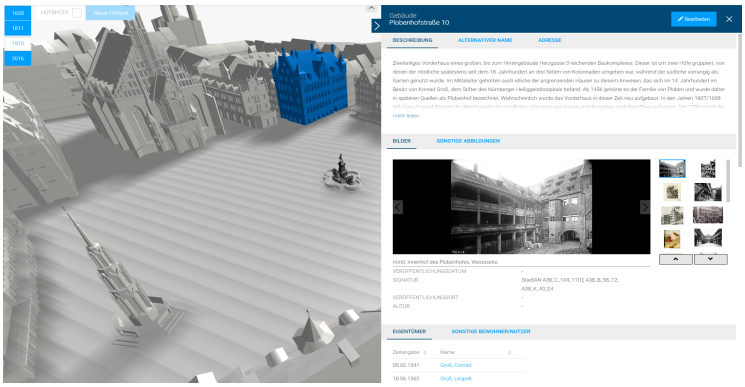

Fig. 1: A screenshot of the city square as represented in the 3D environment (left) and the additional information attached to the selected house (right).

connection of the architectural objects to archival source material along with information on their origin, function and significance. These sources can be historical photographs, drawings, graphics, handwritten sources, address books and chronicles. Linking these information directly to the buildings in the 3D environment enables to explore the way of life in Nuremberg directly at the spatio-temporal point of action. Researchers are then able to explore the architectural changes along with the development of the public life. It will be possible to explore the development of craftmanship, arts, and culture to understand the way technological advancements spread throughout the city and to research the dynamics of local industries. In the predecessor project TOPORAZ, preliminary work on the VRE has been completed and is currently extended with a coverage of more than 3.000 buildings, enriched with historical data from several sources. An example of the exploration environment is pictured in figure 1. The user can explore the 3D space and a click on a building in the model triggers the information attached to a building. The 3D environment along with the connected data will be made available for researchers as well as for educational purposes and the general public. Fundamental for this exploration is the connection of accurate historical data of different time periods to the interactive 3D surface using an intelligent data model. This will be achieved by a knowledge graph which is currently under development. GLAM institutions will have the possibility to connect their resources to the KG provided in this effort which widens the research area and allows to draw cross-connections between repositories. In the following section, the data that has to be curated and linked to the 3D environment to allow sophisticated research will be described, as well as the overall envisioned workflow will be presented.

4 Knowledge Graph based Data Curation

An important step to depict the history of Nuremberg in a 3D VRE is to connect architectural objects to the people who lived there or owned them at a certain

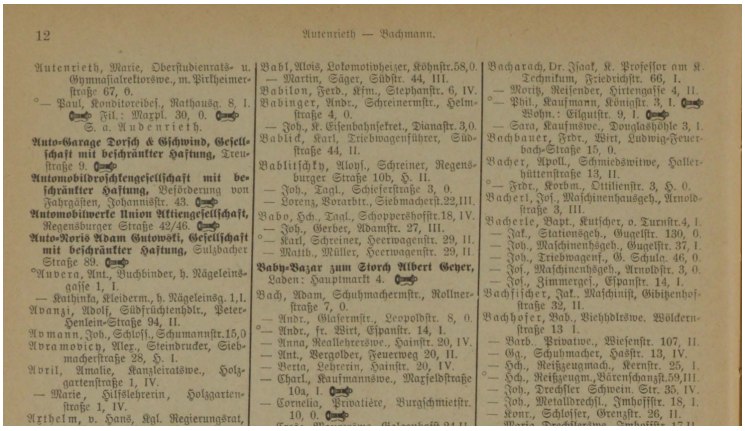

Fig. 2: A digitized page from "Addressbuch von Nürnberg 1910"

.

point in time. This allows researchers to gain a better understanding of the development of the city and its social networks as well as to add depth into a genealogical research of a family.

4.1 Data and Data Sources

In this section, a selection of currently processed data sources is introduced with the goal to integrate the data into the TRANSRAZ KG and to connect it to corresponding objects in the 3D model.

Address Books. One of the first printed sources that contains information about persons living in Nuremberg are the address books, which are physically stored in Nuremberg City Archives[6] and in The Germanisches Nationalmuseum[7]. The annually published books starting with the year 1792 are provided digitally in as scanned images as shown in Figure 2. To access the contained text information, Optical Character Recognition (OCR) has to be performed. Challenges of the transcription process are manyfold and range from bad paper quality, distortion of pages, and poor inking up to exceptional linguistic features. Antiquated fonts, ligatures, archaic terms, old spelling variants, abbreviations and typos are common characteristics of these historical documents that complicate the recognition process. An additional challenge is to capture semantics encoded in non-alphanumeric symbols and specific font features. For the sake of correct text segmentation and in order to avoid producing chaotic text blocks, most of the OCR systems are trained to remove any of such content from the image before the recognition. However, in historical documents such characteristics often contain meaning. For example, people whom civil rights were granted are prefixed with a circle (○) and households with a telephone are denoted with a handset icon.

[6] https://www.nuernberg.de/internet/stadtarchiv_e/
[7] https://www.gnm.de/en/museum/

Nuremberg Artists Lexicon. The "Nürnberger Künstlerlexicon" (NKL) [4] is a collection of bibliographical articles about artists of Nuremberg based on various archival records ranging from the 12th century to the mid 20th century. The articles provide both personal information of artists such as addresses, professions, birth and death places and dates, family relations, places and periods of study, and information about their artworks and their public life. The articles of NKL are based on administrative records, the text is saturated with temporal units to describe the events. However, information is not always provided in complete sentences, most of them are lacking subjects or predicates, thus making the relation extraction more challenging.
MVGN. The "Mitteilungen des Vereins für Geschichte der Stadt Nürnberg"[8] (Journal of the Association for History of the City of Nuremberg) is a journal that publishes scholarly articles on all areas of the history of Nuremberg. Since 1879 an annual issue contains up to 40 reviews on important events, persons and every day life of the citizens. In cooperation with the Association for the History of the City of Nuremberg[9] and the Bavarian State Library[10] the MVNG were scanned and are now available online, however not in a textual form. A defining feature of the articles in the early issues of MVGN is a long compound structure of the sentences filled with a wide range of coordinating conjunctions and descriptive introductory phrases and sentences.
Books of Nuremberg's Twelve Brothers. "Nürnberg Zwölfbruderbücher"[11] were first created in the middle ages as a collection of portraits and biographical data of old Nuremberg craftsmen that decided to retire in an old's people home. The books were previously digitized, transcribed and indexed[12]. During the transcription process no adjustment of the modern spelling was conducted. While indexing the entries, however, an alternative spelling with the modern alphabet and orthography was provided, and the abbreviations were resolved. The data is semi-structured, i.e., every entry contains information on first names, last name, professions and different spelling variations of the data, the birth and death dates and places, the date when the person was registered in the retirement home and the duration of the stay. Also, descriptions of the person's portrait are mentioned.

4.2 Information Extraction Workflow

To enable the exploration of the described historical data by means of the 3D VRE, the scanned documents will be transferred to machine understandable data. The workflow can be divided into four main steps which will be described below:
OCR. Today, OCR systems applied on modern fonts reach such high accuracy that it is considered an almost solved research task. However, such models do not produce equally satisfying results if applied to historical writing. Several

[8] https://www.bayerische-landesbibliothek-online.de/mvgn
[9] https://www.nuernberg.de/internet/stadtarchiv/vgn.html
[10] https://www.bsb-muenchen.de/en/
[11] https://hausbuecher.nuernberg.de/index.php?do=page&mo=2
[12] https://hausbuecher.nuernberg.de/index.php?do=page&mo=5

models are trained to recognise historical text (e.g., Tesseract OCR[13] or Calamari OCR[14]). The address books described above mostly contain proper names and historical job titles. Therefore, common approaches as e.g., high frequency words [8] cannot be applied in a meaningful way. Systematic errors of the OCR, e.g., orthographic errors due to visual letter similarity (*Ste.*" instead of *Str.*(Straße)), segmentation mistakes, etc. can be corrected, and the data can be structured with the help of regular expressions. Spelling mistakes can be resolved by using a reference lexicon and lookup matching of potential candidates in the lexicon for correction, based on the Levenshtein distance. Since the *Books of Nuremberg's Twelve Brothers* contain a manually composed person index, it can be applied as a reference vocabulary for error correction in first and last names. Also, the list of street names represented within the 3D model can be used to correct the street names. For historical job titles external resources can be leveraged, as e.g., the instances of Wikipedia category "Historischer Beruf" (historic profession)[15].

NER. The historical articles contained in NKL and MVGN include a huge amount of dates that describe events in the lives of the persons contained therein, which causes ambiguity, for example, in sentence *"During the Spanish Civil War January 10, 1937 he fled to Switzerland"* the given date may be mistakenly assigned to the war period. To resolve the problem, the text will first be segmented by extracting the dates and date-ranges along with the event associated with the date by using existing NER techniques, e.g., Stanford NER[16] and SpaCy[17]. After the data is segmented, other named entities, e.g., names, locations, professions will be extracted.

Relation extraction. After the named entities are recognized, the relations within the data have to be defined and corresponding triples have to be generated. The sentence structure of articles, e.g., in MVGN, is highly complex and the information provided there is too extensive (i.e., each article has a main subject with unstructured natural language text related to it). Therefore, a set of the most relevant relations is composed together with domain experts. These relations are used to describe birth and death dates, the date of marriages and divorces and the date a person has entered or left a working position. Based on this list, a subset of the textual data will be selected and the extracted named entities will be linked to the subject of the article (e.g., person, organization) via predefined relations. This can be done by leveraging the format of the semi-structured contents, or by exploiting data mining-driven approaches to relate the content of texts to the designed ontology properties for the unstructured contents. Finally, to represent the temporal context of a triple, several methods might be used, e.g., reification [3] or RDF* [5].

Knowledge graph integration. Once the data sources have been analyzed, the extracted information is represented in a KG. A major challenge here is

[13] `https://github.com/tesseract-ocr`
[14] `https://github.com/Calamari-OCR/calamari`
[15] `https://de.wikipedia.org/wiki/Kategorie:Historischer_Beruf`
[16] `https://nlp.stanford.edu/software/CRF-NER.shtml`
[17] `https://spacy.io/usage/linguistic-features#named-entities`

the accurate alignment of entities (people, organizations, events). Furthermore, the extracted data have to be integrated into the overall platform environment including the interactive 3D model. As part of the predecessor project TOPORAZ, a relational database was created linking data with the 3D model. However, this approach lacks interoperability and, therefore, an ontology is currently under development to describe and represent the historical city architecture, and connect the resources described in 4.1 to the 3D model. Existing ontologies and thesauri are reused for this effort, including *CIDOC CRM*[18], *ArCo* [2] and the controlled *Art and Architecture Thesaurus*[19].

4.3 Data Curation Challenges

In this section, the project's main challenges are presented, how they may be tackled, and how the expected results might support the understanding and curation of historical data.

Information extraction. TRANSRAZ builds on heterogeneous data sources that contain information that can be parsed, structured, and linked for a better curation. However, there are challenges that need to be addressed to correctly integrate them into the exploration environment. To start with, data sources describe Nuremberg's citizens without providing unique identifiers to distinguish them. For example, people who share the same name might be considered the same person, generating incorrect information about the story of the Nuremberg city. To tackle this issue, automated methodologies based on local information about people, places, and events will be studied to provide certainty of the correctness of the knowledge graph links.

Language evolution and meaning shift. Another important challenge that needs to be addressed when parsing historical data is the evolution of the language. Obsolete words that are not used anymore, or words that changed their meaning, can lead to grasp an incorrect message from the sources of data and, therefore, to generate knowledge graphs that can mislead the correct interpretation of historical facts. An example is represented by the evolution of names and types of professions throughout the centuries.

Temporal component. Understanding and representing how the interactions between people, places, and events evolved in the city of Nuremberg require the association of temporal components. In fact, there might be relations between historical entities that existed only in a specific point in time or intervals (for example a person who lived in the city from 1811). This kind of information can be easily interpreted by humans, but it might become challenging when formal representations are required to make the information machine understandable. Therefore, one important goal of the project is to come up with a reasonable representation of time and temporal relations by means of ontologies to provide appropriate means for correct interpretation of the represented facts.

[18] `http://www.cidoc-crm.org/`

[19] `https://www.getty.edu/research/tools/vocabularies/aat/`

5 Conclusion

In this position paper, the vision of the project TRANSRAZ is explained in which KGs are created to connect heterogeneous historical data about people, organizations and events in the historic city of Nuremberg to an interactive 3D model. The nature of the data sources provide a number of curation challenges which are introduced in the paper. Most of these historical data are currently only available as scans and are hidden in archives unable to be explored by anyone. Publishing the data using KGs allows to make these sources available for an intuitive exploration on the Web, and enables to connect these resources to further repositories by GLAM institutions for research and education.

Acknowledgement. We would like to thank Christiane Stöckert, Felix Schönrock and Gerhard Weilandt for their indispensable input as domain experts. This work is funded by the Leibniz Association under project number SAW-2020-FIZ KA-4-Transraz.

References

1. Bukhari, S.S., Kadi, A., Jouneh, M.A., Mir, F.M., Dengel, A.: anyocr: An open-source ocr system for historical archives. In: 2017 14th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR). vol. 1, pp. 305–310. IEEE (2017)
2. Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., Veninata, C.: Arco: The italian cultural heritage knowledge graph. In: Int. Semantic Web Conference. pp. 36–52. Springer (2019)
3. Cyganiak, R., Wood, D., Lanthaler, M.: Rdf 1.1 concepts and abstract syntax (2014)
4. Grieb, M.H.: Nürnberger Künstlerlexikon: Bildende Künstler, Kunsthandwerker, Gelehrte, Sammler, Kulturschaffende und Mäzene vom 12. bis zur Mitte des 20. Jahrhunderts. Walter de Gruyter (2011)
5. Hartig, O.: Foundations of rdf* and sparql*:(an alternative approach to statement-level metadata in rdf). In: AMW 2017 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web, Montevideo, Uruguay, June 7-9, 2017. vol. 1912. Juan Reutter, Divesh Srivastava (2017)
6. Otten, H., Hildebrand, L., Nagel, T., Dörk, M., Müller, B.: Shifted maps: Revealing spatio-temporal topologies in movement data. In: IEEE VIS Arts Program. pp. 1–10. IEEE (2018)
7. Razum, M., Göller, S., Sack, H., Tietz, T., Vsesviatska, O., Weilandt, G., Grellert, M., Scharm, T.: Toporaz: Ein digitales raum-zeit-modell für vernetzte forschung am beispiel nürnberg. Information-Wissenschaft & Praxis **71**(4), 185–194 (2020)
8. Reynaert, M.: Non-interactive OCR post-correction for giga-scale digitization projects. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 617–630. Springer (2008)
9. Tietz, T., Waitelonis, J., Zhou, K., Felgentreff, P., Meyer, N., Weber, A., Sack, H.: Linked stage graph. In: SEMANTICS Posters&Demos (2019)
10. Tolstoi, P.: A framework for location-based augmented reality content on mobile devices (2019)
11. Vsesviatska, O., Tietz, T., Hoppe, F., Sprau, M., Meyer, N., Dessi, D., Sack, H.: Ardo: An ontology to describe the dynamics of multimedia archival records [to be published]. In: ACM, Symposium On Applied Computing (2021)