Towards a Representation of Temporal Data in Archival Records: Use Cases and Requirements

Oleksandra Bruns[1,2], Tabea Tietz[1,2], Mahsa Vafaie[1,2], Danilo Dessí[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. Archival records are essential sources of information for historians and digital humanists to understand history. For modern information systems they are often analysed and integrated into Knowledge Graphs for better access, interoperability and re-use. However, due to restrictions of the representation of RDF predicates temporal data within archival records is a challenge to model. This position paper explains requirements for modeling temporal data in archival records based on running research projects in which archival records are analysed and integrated in Knowledge Graphs for research and exploration.

Keywords: Knowledge Graphs · Archives · Temporal Representation.

1 Introduction

The research, exploration, and analysis of historical events throughout different time layers are fundamental for understanding their influence on today's world as well as for increasing the sense of cultural identity. Being primary sources of information, archival records play an essential role in providing real evidence on events. Therefore, a structured and formal representation of millions of existing archival resources from different time periods is necessary to preserve, to explore and to enrich the timeline of history, e.g., What event triggered a certain decision? How did a city develop over a certain period of time? How did legal processes and court proceedings change over time? What terminology was used several centuries ago and how did it evolve? In order to provide novel means of exploration, a promising approach would be the integration of heterogeneous archival records from various points in history into semantically interlinked representations such as Knowledge Graphs (KGs). For this purpose, the time varying data and knowledge contained in such records must be represented in a semantically correct way and furthermore reasoned and queried. Despite the high amount of research ideas for integrating the concept of time into Semantic Web applications (cf. [15] for a corresponding survey), the proposed approaches differ from each other in various ways, and thus the choice of a certain modeling approach is highly dependent on its respective context. For example, the approaches differ in the dimensions of time they adapt: whether it is the valid time to represent the validity of a

certain fact, concept or event [6,8,9,11], or the transaction time to record the time of change in a KG [3]. They differ on the type of temporal data, whether only temporal change of facts is considered [5,8,9], or also temporal existence of entities or concepts [6,16] has to be denoted. Furthermore, the temporal statements in data can be expressed differently. An additional challenge lies in the nature of historical records where the information about the validity of a fact or a concept is uncertain or partly or fully missing, therefore the documentation (and thereby representation) of a complete provenance is essential.

The contribution of this position paper is the presentation of use cases which utilize archival records for research purposes, the description of the nature of temporal data in them, and the discussion of resulting requirements for KGs to enable modeling and reasoning over temporal data in archival records.

2 Use Cases: Archival Data for Historical Research

In the following, three use cases are introduced. They form the basis of discussion to identify the requirements for the representation of temporal data in archival resources.

2.1 Use Case 1: Exploring Nuremberg in different Time Layers

Within the TRANSRAZ project, a KG based 3D virtual research environment (VRE) is created to explore the city of Nuremberg in different time layers from the middle ages to modern times. The VRE connects buildings with information about historical places, events, residents and organizations extracted from external resources [10,13]. Exemplary heterogeneous data sources relevant for the discussion on temporal representations of archival data are presented in the following.

Use Case 1.1: Historical Address Books. Nuremberg address books provide a valuable resource to explore the city history. They have been published since 1792 and provide information on persons, their occupations, addresses as well as companies. The address book data from different time layers is in the process of being integrated into a KG to explore [1]. A major challenge of the integration is the disambiguation and alignment of entities over a period of time, since all information is dynamic. Persons change their names, addresses and occupations. Streets are renamed, building numbers change and buildings are destroyed and rebuilt. Occupations may either change their name or their function.

Use Case 1.2: The "Nürnberger Künstlerlexicon" (NKL). NKL [4] is a collection of bibliographical articles about artists of Nuremberg based on archival records ranging from the 12th to the 20th century. The articles provide both personal information of artists such as occupations, 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, and the text is saturated with varying temporal units to describe events.

A challenge in the process of integrating this information into a KG is the description of events and dates, which includes time instants (e.g. September 27th, 1763) and time intervals (e.g., from 1763 to 1782), but also unknown time stamps (e.g., until his death), incomplete intervals (e.g. from 1756) and temporal uncertainty (e.g. around 1830).

Without a semantically correct representation of these dynamics in the TRANSRAZ KG, cultural heritage researchers cannot be supported in tasks involving time dependent data exploration, e.g. When did important artists settle in the city, where did they live and when did they move to different city areas? Was their movement triggered by significant events in history?

2.2 Use Case 2: Archivportal-D: Subject-related Points of Access

With the nationwide Archivportal-D, a platform is created to make institutional information, indexing services and digitized archival holdings available to everyone for scientific use through a central point of access. The first subject-related information that can be accessed via Archivportal-D is holdings and data related to the Weimar Republic[3]. To collect historical records, store them long-term and make them accessible for everyone, archival resources have to be categorized to provide a structured and intuitive access for search and exploration. For this purpose, new classification schemes have to be created for subject specific entries depending on the topic of the archival records. Furthermore, whenever another archive provides access to additional records within the same subject specific platform, and whenever new records are digitized, a classification scheme has to be accordingly adapted, i.e. it is dynamic. When creating an ontology that models the classification of these topic based archival resources, these dynamics have to be specifically addressed [7,14]. In the end, researchers can draw conclusions on the document classification process, e.g. Which keywords were eliminated and which documents are more controversially addressed and were re-classified the most? Another challenge is the representation of the time periods the documents describe. Also in this use case, vague time intervals and uncertainty have to be represented within a KG as precise as possible from a historical perspective and semantically correct for data interoperability, data exploration and semantic reasoning. This allows to conclude on the timeline of events the records describe, e.g. What event or action resulted from a certain political statement?

2.3 Use Case 3: Wiedergutmachung

"Pilotprojekt zur Wiedergutmachung"[4] is issued by the German Federal Ministry of Finance. It is centred around archival data from the reparation process and reparation cases filed after World War II and the fall of the Nazi regime, in the state of Baden-Württemberg. *Wiedergutmachung* includes the reparation, restitution, and indemnification for victims of Nazi persecution [2]. The goal

[3] `https://www.archivportal-d.de/themenportale/weimarer-republik`

[4] `https://www.fiz-karlsruhe.de/en/forschung/wiedergutmachung`

of this project is to integrate archival data into a KG enabling its enrichment through connection to external resources. Every archival record set corresponds to a court case which starts after the application is sent to the office of reparation. The record sets are closed shortly after the court case is concluded and the decision about the entitlement to and amount of reparation is made. The archival metadata provides information on the "duration of file" for each record set. However, the duration of the legal procedures at the court remains unknown. For modeling interval relations between the court cases and the archival record sets documenting them, along with interval relations between different court cases, a data model that enables anonymous timestamps is required. Another challenge in modeling the reparation cases arises from the fact that official documents usually have more than one date associated with them. In the case of Wiedergutmachung documents, date of issue in most cases differs from date of processing by the responsible state office. Thus, to reproduce the timeline of the court case it is essential for such data to be extracted and represented in a KG for exploration. Furthermore, some of the documents and forms changed over time from the late 1940s to the early 1960s. Such diachronic changes in the content of the resources can be captured by a temporal data model and integrated into a KG.

3 Discussion of Requirements

This section presents a preliminary analysis of the historical records discussed in section 2. The following modeling requirements (REQ) were identified:

REQ1: Modeling change – Relabeling and Concept Drift. Since archival records cover different periods of time, it is natural that mentioned concepts and entities change. Such changes can happen on the name level, when the meaning stays the same and only a new name for the concept or entity is provided. In address books from different years (use case 1.1) the same person may have different last names due to marriage, the street may be renamed due to historical events, the occupation names may be outdated and substituted with more modern terms, e.g., "mannequin" is an outdated term for "model". In use case 2, labels of concepts in classification schemes are often adapted by archivists to provide a more suitable representation, e.g. the keyword "May Day" was later renamed into "Labour Day". In use case 3, both personal information such as person names and names of the state offices may differ throughout the time.

Alternatively, the meaning of some concepts evolves without changing the name [12]. A typical example of such changes is occupation names that occur in use cases 1 and 3. Occupation functions may change over time, so their definition will vary across different time layers, even though the name stays the same. In use case 2 the change of meaning is observed whenever the classification scheme is adapted. For example, the category "Road Traffic" would have a broader meaning with the addition of the subcategory "Bicycle" to the classification scheme.

REQ2: Modeling time dependent information – Temporal Facts, Concepts, Entities. Temporal data in archival records varies and is associated with

both point-based and interval-based timestamps. In use case 1.1, only one type of time units is provided – issue year of the book. Thus, every fact has to be associated with a year interval. Use cases 1.2 and 3 contain both instant and lasting facts, and have to be annotated with either timestamps or time intervals, e.g., *"Johann Ackermann married Elisabatha Schönberger on November 20, 1834"* and *"Johann Ackermann studied medicine in Jena and Göttingen from 1668 until 1771"*. Similar to facts, concepts and entities may be denoted with their existence in time. For use cases 1 and 3 the existence of occupations has to be annotated, e.g. the profession "knocker-up" existed only until the 1940s. For use case 1.1 buildings have to be associated with their construction and destruction dates. In use case 3, types of documents and forms changed overtime and may be denoted with their period of use. In use case 2, every change in the classification scheme is associated with the new version of the scheme. Hence, this version entity is to be annotated with the time interval it was or is valid in.

REQ3: Modeling timestamps – Anonymous and Uncertain Time. Due to their nature, historical records often contain incomplete temporal information that must be modeled via anonymous timestamps (variables). For all use cases, it is often the case when facts or entities lack information about their validity, however cannot be considered static, e.g., address of a person. In some cases temporal intervals are not complete, e.g. in a sentence *"Reinhold Bach worked in Nuremberg until 1923"* the beginning of the interval is unknown. However, modeling of anonymous time annotations is important for stating the interval relations between facts. Finally, some archival records from all use cases do not contain information on an original date of issue, but they contain subsequent annotations of approximate dates by historians, e.g. *"the beginning of the 1950s", "23.08.1908 or 24.08.1908", "around 1872"*. Such timestamps cannot be considered anonymous and have to be normalized for reasoning.

In summary, archival resources are heterogeneous and contain diverse temporal information. The representation of time varying data in archival records is complex and requires a highly expressive approach for the enrichment and exploration of our past stored in archives.

4 Conclusion

In this position paper, use cases that utilize archival records for research purposes are presented. Furthermore, requirements for modeling temporal data within historical archival records are defined. In future work, more use cases will be employed to provide a more complete and better structured list of requirements to develop guidelines for proper temporal semantic annotation of archival records. Based on the guidelines, a semantically correct and historically accurate representation of temporal data will be published to enable federated time-based queries over different archives to provide researchers with a holistic view into our past.

Acknowledgement. This work is funded by the Leibniz Association under project number SAW-2020-FIZ KA-4-Transraz and by the Deutsche Forschungsgemeinschaft with grant number 396707386.

References

1. Bruns, O., Tietz, T., Chaabane, M.B., Portz, M., Xiong, F., Sack, H.: The Nuremberg Address Knowledge Graph. In: 18th Extended Semantic Web Conference (ESWC), Poster and Demo Track. Springer (2021)
2. Goschler, C.: The United States and Wiedergutmachung for Victims of Nazi Persecution: From Leadership to Disengagement. Holocaust and Shilumim: The Policy of Wiedergutmachung in the Early 1950s. Washington, DC: German Historical Institute (1991)
3. Grandi, F.: Multi-temporal RDF Ontology Versioning. In: Proceedings of the 3rd International Workshop on Ontology Dynamics (IWOD-09). pp. 625–634 (2009)
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. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Artificial Intelligence **194**, 28–61 (2013)
7. Hoppe, F., Tietz, T., Dessı, D., Meyer, N., Sprau, M., Alam, M., Sack, H.: The Challenges of German Archival Document Categorization on Insufficient Labeled Data. In: WHiSe 2020: Workshop on Humanities in the Semantic Web 2020-Proceedings of the Third Workshop on Humanities in the Semantic Web (WHiSe 2020), co-located with 15th Extended Semantic Web Conference (ESWC 2020), Heraklion, Greece, June 2, 2020 (online). Ed.: A. Adamou (2020)
8. Hurtado, C., Vaisman, A.: Reasoning with Temporal Constraints in RDF. In: International Workshop on Principles and Practice of Semantic Web Reasoning. pp. 164–178. Springer (2006)
9. Nguyen, V., Bodenreider, O., Sheth, A.: Don't like RDF reification? Making statements about statements using singleton property. In: Proceedings of the 23rd international conference on World wide web. pp. 759–770 (2014)
10. 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)
11. Schueler, B., Sizov, S., Staab, S., Tran, D.T.: Querying for Meta Knowledge. In: Proceedings of the 17th international conference on World Wide Web. pp. 625–634 (2008)
12. Tietz, T., Alam, M., Sack, H., van Erp, M.: Challenges of Knowledge Graph Evolution from an NLP Perspective. In: WHiSe 2020: Workshop on Humanities in the Semantic Web 2020-Proceedings of the Third Workshop on Humanities in the Semantic Web (WHiSe 2020), co-located with 15th Extended Semantic Web Conference (ESWC 2020), Heraklion, Greece, June 2, 2020 (online). Ed.: A. Adamou (2020)
13. Tietz, T., Bruns, O., Göller, S., Razum, M., Dessì, D., Sack, H.: Knowledge graph enabled Curation and Exploration of Nuremberg's City Heritage. In: Proceedings of the Conference on Digital Curation Technologies (Qurator 2021): Berlin, Germany, February 8th to 12th, 2021. Ed.: A. Paschke (2021)

14. 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. In: Proceedings of the 36th Annual ACM Symposium on Applied Computing. pp. 1855–1863 (2021)
15. Wang, H.T., Tansel, A.U.: Temporal Extensions to RDF. Journal of Web Engineering **18**(1), 125–168 (2019)
16. Welty, C., Fikes, R., Makarios, S.: A Reusable Ontology for Fluents in OWL. In: FOIS. vol. 150, pp. 226–236 (2006)