Knowledge Graph Construction and Refinement for Cultural Heritage Digital Libraries

Mary Ann Tan[1,2]

[1]*FIZ Karlsruhe – Leibniz Institute for Information Infrastructure,*
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

[2]*Applied Informatics and Formal Description Methods (AIFB), Karlsruhe Institute of Technology (KIT),*
Kaiserstraße 89, 76133 Karlsruhe, Germany

Abstract

Digital Libraries containing metadata of diverse cultural heritage objects are meant to be accessible not only to domain experts but also to the general population. This calls for information services that can provide ease and efficiency to search, retrieval and exploration. Knowledge graphs (KGs) are essential for representation, organization, integration, and analysis of hierarchical and heterogeneous information. However, most KGs suffer from incompleteness and inaccuracies. This work intends to address various challenges arising from construction and refinement of a KG populated with historical objects, by defining domain- and application-appropriate ontologies and leveraging approaches in information extraction (IE) for improving metadata quality.

Keywords

Semantic Web, NLP, Information Extraction, Knowledge Graphs, Digital Libraries, Cultural Heritage

1. Introduction

The German Digital Library[1] (DDB) collects, aggregates, transforms, and publishes metadata representing tens of millions of digitized cultural heritage objects (eg. books, paintings, archival documents, photographs, audio recordings). These objects span several millennia and belong to the holdings of various memory institutions all across Germany. Due to its historical significance, this collection is meant to be accessed and explored by users from diverse backgrounds.

However, the sheer volume, granularity, and heterogeneity of this collection hampers the ease in search, retrieval, and exploration. These hurdles call for the construction of a knowledge graph (KG) to represent and to organize the objects and their contextual descriptions, while enabling data integration and analytics.

As the national aggregator to the Europeana[1], DDB's metadata collection is represented using an extension of the Europeana Data Model (EDM)[2]. EDM favors simplicity and offers flexibility in the choice of metadata element sets, as well as the range of possible values for properties describing the objects. These design considerations lead to modeling challenges

ISWC 2024 Doctoral Consortium, co-located with the 23rd International Semantic Web Conference, November 11, 2024, Baltimore, Maryland, USA

ann.tan@fiz-karlsruhe.de (M. A. Tan)
0000-0003-3634-3550 (M. A. Tan)

[1] *Deutsche Digitale Bibliothek*, https://www.deutsche-digitale-bibliothek.de

[2] EDM, https://pro.europeana.eu/page/edm-documentation

described by Tan et al. [2]. In addition, the metadata collection suffers from incompleteness and inaccuracies as described in Tan et al. [3]. This prevents the underlying retrieval engine from properly indexing the objects.

To address these challenges necessitates a combination of solutions in knowledge representation, knowledge refinement, and information extraction. Therefore, this thesis proposes i) an ontology that enables interoperability across different types of CHOs while maintaining domain-specific semantics as discussed in Section 5.1; ii) a KG refinement approach leveraging NLP teachniques to improve metadata quality of historical objects; and iii) an Entity Linking approach for entities in historical objects.

2. Importance

This work will benefit not only the general population, but also the domain experts such as librarians, curators, and archivists. Proposed solutions will empower users from diverse backgrounds to seamlessly and efficiently search, retrieve, and explore Germany's rich and voluminous collection.

Recent developments in AI can be leveraged to address the technical challenges facing the DDB. This work is relevant to the researchers working at the intersection of Semantic Web (SW), Digital Humanities (DH), and Natural Language Processing (NLP).

3. Related Work

There have been several notable data models or ontologies proposed for cultural heritage representation. Liu et al. [4] provided a review of CIDOC-CRM, Sampo Model, and EDM specific to the museum use case only. Cultural heritage data models are delineated along two modeling paradigms: object-centric and event-centric. CIDOC-CRM follows the former, while EDM follows a mixture of both paradigms. Object-centric modeling defines attributes directly by describing the object, while event-centric modeling defines these attributes through a series of events associated with the object. Object-centric modeling favors conciseness, while event-centric modeling emphasizes completeness.

A pioneer in the application of of SW technologies, the *Sampo* series of semantic portals showcase the national heritage of Finland. These systems make use of the modular FinnONTO ontology infrastructure [5]. However, FinnONTO is not a full-featured ontology, but a taxonomy of CHOs encoded as Simple Knowledge Organization System (SKOS) concepts. Following the modular modeling approach is Italy's ArCO[3] [6], where each module is intended to describe a CHO[4] in the context of cataloging activities and events.

The core design principles of EDM, and by extension DDB-EDM, lead to definitions of general classes that require the bare minimum of metadata properties and controlled vocabularies. Thus, all CHOs, regardless of their sources, media types and object types, are instances of the class *edm:ProvidedCHO*, while their digitized representations on the Web are instances of the class *edm:WebResource*. This flexibility however results in imprecise representations and loss

[3] ArCO, https://w3id.org/arco

[4] CHO is referred to as "Cultural Property".

of semantics inherent in the original objects [7]. In particular, it is not possible to model the concepts and level of abstractions widely-accepted in the bibliographical domain.

The International Federation of Library Associations and Institutions (IFLA) developed the Functional Requirements for Bibliographic Records (FRBR) [8], where a book can be represented as several entities and the relationships that exist among these entities. A copy of a book (`frbr:Item`) is a specimen or exemplification of a specific publication (`frbr:Manifestation`), which is an embodiment of an expression `frbr:Expression` that realizes the ideas of a creative work (`frbr:Work`).

Most Europeana users are less likely to search for specific items (11.3%) and are more inclined to search by category (47.1%) and by subject (24.6%) [9]. This supports the need to align bibliographic objects from the *Item* level to their respective higher-level abstractions (*Work*, *Expression*, *Manifestation*). Consequently, the process of alignment sets a prerequisite for objects to possess identifiable properties and attributes, such as title, agents, dates, and subject heading. However, due to the age of the objects, a high level of uncertainty with respect to proper author or date attributions is apparent.

The challenges of filling missing information and identifying erroneous information in a knowledge graph fall under the umbrella of Knowledge Graph Refinement. In particular, *Knowledge Graph Completion* (KGC) deals with the former challenge, while *Error Detection* deals with the latter.

By definition, internal methods for KGC use the content of the current KG either to determine class membership or to predict relations between entities. These methods require the current KG to at least possess reasonable quality in order for large scale evaluation to be feasible [10]. On the other hand, external methods leverage other sources of knowledge for refinement, such as other knowledge graphs or text corpora.

With the rapid development in the area of Natural Language Processing (NLP), text corpora have become an excellent source of external knowledge. The subfield of Information Extraction (IE), an intermediate step to knowledge graph construction, can be defined as the process of gleaning structured information from unstructured text [11]. A concrete example of this task would be to extract distinct properties and attributes identifying a literary work from the title.

An IE pipeline starts with Named Entity Recognition (NER), or the detection and classification of named entities mentioned in the text. Types of entities can be coarse-grained such as `PERSON`, `WORK_OF_ART`, `DATE`, et cetera or fine-grained such as `AUTHOR`, `PUBLISHER`, `ARTWORK`, `PUBLISHER`, `LITERARY_WORK`, `PUBLICATION DATE`, et cetera.

Specific entity types (fine-grained) are often found in domain-specific texts, or even time-specific texts where concept drift is quite common. In the field of digial humanities, there are a number of studies on NER with historical text [12], however, coverage goes back to 17 th century B.C. at best (DROC [13]) and none belong to the domain of bibliography.

Once the entities have been detected and classified, they are linked to specific entries in reference knowledge bases or KGs. Entity Linking is particularly challenging due to the surface form variations. In particular, names in historical texts can be multilingual, refer to aliases or contain initials, include honorifics and designations. The names of geop-olitical entities are also known to change through time. Pontes et al. [14] proposed an end-to-end multilingual NER and EL (NERL) approach to address some of these challenges using some of the datasets mentioned in Ehrmann et al. [12].

4. Research Questions

This section formulates the research questions (RQs) to address the challenges and limitations of existing approaches described in Section 3.

RQ1: How can existing ontologies be adapted and extended to suit the domain and application profile of digital libraries, such as the DDB?
Cultural heritage practitioners have been developing ontologies for specifc domains and applications. As of this writing, only EDM is used to represent metadata from several cultural institutions. In order to prevent data model silos [15] and to promote reusability, it is beneficial to consider existing ontologies that are applicable and appropriate for the use case of the DDB. Preliminary results are discussed in Section 5.1.

RQ2: How can we leverage state-of-the-art NLP models to improve metadata quality of historical objects?
Non-contemporary titles in the DDB (`<dc:title>`) encode details that can be used to fill-out missing properties, such as the title itself, author, publisher, editor, subject headings, and dates. Hence, this calls for extractive NLP approaches. Section 5.2 presents some preliminary results.

RQ2.1: How can we automatically construct an evaluation dataset from the DDB?
In order to address the succeeding RQs, an evaluation dataset for IE is required. Section 5.2 briefly describes what has been done so far.

RQ2.2: How can we effectively extract fine-grained bibliographic entities from historical texts?
The goal here is to address open challenges in the area of historical NER, such as how to properly handle the dynamics of an evolving language, where spelling and naming conventions change through time, and noise resulting from OCR engine. Dataset construction, design of experiments, and model development will be accomplished.

RQ3: How can we link entities to records in the reference KG? The goal here is to accurately disambiguate named entities and link them to entries in external KGs, while addressing the challenges associated with historical texts. Moreover, entities that do not exist in the reference KG can used as further contribution to increase the coverage of authority files.

The entirety of this work is envisioned to guide the construction and refinement of a knowledge graph representing DDB's cultural heritage objects. In addition, some open questions have yet to be addressed concerning NERL in historical texts.

5. Preliminary Results

The section describes preliminary work conducted to address the open questions presented in Section 4.

5.1. The DDB Ontology (DDB-O)

Extensive quantitative and qualitative analysis of the entire DDB metadata collection have been conducted in order to ascertain the applicability of existing CH ontologies. Initially, objects were logically classified according to their originating institution, whether from libraries, archives, museums, media libraries, or historical preservation. In addition, the media type of an object was also taken into account. Taking up a large proportion of the entire collection, the alignment of textual bibliographic resources to an extension of FRBR [5] have been presented [2] and implemented as a SPARQL Endpoint [16]. Domain-specific ontologies have been adapted to have more precise semantic representation objects (eg. components of bibliographic objects, hierarchy of archival objects, level of representations of an image, etc.) Existing audio ontologies intended for other domains have been extended to represent intangible audio heritage [17]. The DDB-O Namespace[6] is available online. A formal and complete specification is under review and yet to be published.

FRBR, as the upper ontology, requires that each object is looked up against a list of creative works, such as the German Authority File or *Gemeinsame Normdatei* (GND[7]). This ensures that the relationship between different objects resulting from the same creative work is represented in the KG [18].

5.2. Information Extraction

The alignment of bibliographic items to their corresponding literary works proved to be a challenging task due to incomplete object descriptions [18]. Taking advantage of the greater textual content encoded in the titles, several NLP tasks were reformulated in order to extract contextual details present in the title. Several state-of-the-art, off-the-shelf NER and extractive QA models, as well as LLMs were used in the experiments.

As described in [3], the objects in the evaluation dataset were selected according to language, hierarchy type, existence of agent and date properties, format, and title length (>30 tokens).

A more forgiving evaluation measure (*Precision@n*) described in Section 6.2 was defined to take into account the various naming conventions found in the text. An NER model (FLERT) [19] that can detect literary works and dates was initially used to test the hypothesis, and to refine the evaluation dataset for the succeeding tasks. The results shown in Table 1 illustrate that these models can be leveraged but only to a lesser extent. The results were poor since the models were not adapted to the age and domain of the texts. In addition, the results are not indicative of the actual model performance due to evaluation dataset inaccuracies [3].

6. Evaluation

The research questions enumerated in Section 4 require different evaluation procedures, dataset, and metrics. These are described in the succeeding subsections.

[5] Functional Requirements for Bibliographic Records [8]

[6] DDB-O , https://ise-fizkarlsruhe.github.io/ddbkg/ddbo

[7] GND, https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html

Table 1
LLM vs Extractive QA

Question: ”Who is the ...?	Ground Truth	**LLM** `mistral-7b-instruct-v0.2`	**QA** `gelectra-large-germanquad`
Author	all agents	51.60%	**66.23%**
	`<dc:creator>`	**37.60%**	32.19%
Publisher	`<dc:publisher>`	**2.70%**	0.85%

6.1. Ontology Evaluation

There several ways in evaluating ontologies. One of which is using is using competency questions (CQs). A collection of CQs published in GitHub[8] are included in the partial ontological definitions and alignment activities. In addition, SPARQL query processing time for CQs that can be answered with DDB-EDM will be compared with queries using the proposed ontology.

6.2. Information Extraction

Name matching for historical documents is non-trivial due to various naming conventions and spelling variations. In a QA task, the most forgiving measure is *Accuracy@1*, which returns 1 if there is a single token overlap between the ground truth and the answer. *Precision@n* measure is a combination of 2 matching criteria: an exact match of the DDB object ID and an approximate match for names using the Levenshtein edit distance [3].

The evaluation measures for RQ3 will not be any different from those associated with EL. A large proportion of the agents in the DDB are already linked to GND Persons. And there already exist links between GND and Wikidata entities. This means that it is trivial to combine naming variations and multilingual names for the evaluation dataset. Evaluating geopolitical entities will require prior knowledge of the age of the object in question.

7. Limitations and Future Work

As discussed in Section 5.2, the lack of a gold standard evaluation dataset brings a level of uncertainty to the experimental results. This will be addressed with the creation of a manually annotated dataset with fine-grained entities. Consequently, this dataset will be used to address RQ2.2. In addition, the work conducted to address RQ1 need to be finalized. Finally, entities that already exist in GND will be linked, while non-existing ones can be used to further increase the coverage of GND and Wikidata.

Acknowledgments

I would like to thank my supervisors Prof. Dr. Harald Sack and Dr. Shufan Jiang for their invaluable mentoring and support.

[8] CQs for DDB-O, https://ise-fizkarlsruhe.github.io/ddbkg/docs/examples/

References

[1] J. Purday, Think culture: Europeana.eu from concept to construction, Bibliothek Forschung und Praxis 33 (2009) 170–180. doi:10.1515/bfup.2009.018.

[2] M. A. Tan, T. Tietz, O. Bruns, J. Oppenlaender, D. Dessì, H. Sack, DDB-EDM to FaBiO: The Case of the German Digital Library, in: Proc. of the 20th Int. Semantic Web Conference - Posters and Demos – ISWC 2021, volume 2980, CEUR-WS.org, 2021.

[3] M. A. Tan, S. Jiang, H. Sack, Great Article, in: Workshop on Deep Learning and Linguistic Linked Data, 2024.

[4] F. Liu, J. Hindmarch, M. Hess, A review of the cultural heritage linked open data ontologies and models, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-M-2-2023 (2023) 943–950. URL: https://isprs-archives.copernicus.org/articles/XLVIII-M-2-2023/943/2023/. doi:10.5194/isprs-archives-XLVIII-M-2-2023-943-2023.

[5] E. Hyvönen, K. Viljanen, J. Tuominen, K. Seppälä, Building a national semantic web ontology and ontology service infrastructure –the finnonto approach, in: S. Bechhofer, M. Hauswirth, J. Hoffmann, M. Koubarakis (Eds.), The Semantic Web: Research and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 95–109.

[6] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, C. Veninata, ArCo: The Italian Cultural Heritage Knowledge Graph, The Semantic Web – ISWC 2019 (2019) 36––52. doi:10.1007/978-3-030-30796-7_3.

[7] S. Peroni, F. Tomasi, F. Vitali, Reflecting on the Europeana Data Model, in: IRCDL 2012, 2012, pp. 228–240.

[8] B. Tillet, What is FRBR?: A Conceptual Model for the Bibliographic Universe, 2004.

[9] P. Clough, T. Hill, M. L. Paramita, P. Goodale, Europeana: What users search for and why, in: J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, I. Karydis (Eds.), Research and Advanced Technology for Digital Libraries, Springer International Publishing, Cham, 2017, pp. 207–219.

[10] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, Semant. Web 8 (2017) 489–508. URL: https://doi.org/10.3233/SW-160218. doi:10.3233/SW-160218.

[11] M. E. Okurowski, Information extraction overview, in: TIPSTER TEXT PROGRAM: PHASE I: Proceedings of a Workshop held at Fredricksburg, Virginia, September 19-23, 1993, Association for Computational Linguistics, Fredericksburg, Virginia, USA, 1993, pp. 117–121. URL: https://aclanthology.org/X93-1012. doi:10.3115/1119149.1119164.

[12] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet, Named entity recognition and classification in historical documents: A survey, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3604931. doi:10.1145/3604931.

[13] M. Krug, L. Weimer, I. Reger, L. Macharowsky, S. Feldhaus, F. Puppe, F. Jannidis, Description of a corpus of character references in german novels-droc [deutsches roman corpus], DARIAH-DE Working Papers 27 (2018) 1–16.

[14] E. L. Pontes, L. A. Cabrera-Diego, J. G. Moreno, E. Boros, E. L. Pontes, A. Hamdi, N. Sidère, M. Coustaty, A. Doucet, Entity Linking for Historical Documents: Challenges and Solutions, in: 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020, volume

12504 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 215–231. URL: https://hal.science/hal-03034492. doi:10.1007/978-3-030-64452-9_19.

[15] O. Suominen, N. Hyvönen, From MARC Silos to Linked Data Silos, URL: https://swib.org/swib16/slides/suominen_silos.pdf, 2016.

[16] M. A. Tan, T. Tietz, O. Bruns, J. Oppenlaender, D. Dessì, H. Sack, DDB-KG: The German Bibliographic Heritage in a Knowledge Graph, in: 6th Int. Workshop on Computational History at JCDL – Histoinformatics, volume 2981, CEUR-WS.org, 2021.

[17] M. A. Tan, E. Posthumus, H. Sack, Audio Ontologies for Intangible Cultural Heritage, in: Proc. of the 19th European Semantic Web Conference - Posters and Demos – ESWC 2022, 2022.

[18] M. A. Tan, H. Sack, The DDB Collection and the Limits of Artificial Intelligence, URL: https://swib.org/swib23/slides/06_Mary%20Ann%20Tan_SWIB2023%20Final.pdf, 2023.

[19] S. Schweter, A. Akbik, Flert: Document-level features for named entity recognition, 2020. arXiv:2011.06993.