Domain-Specific Knowledge Graph Construction for Semantic Analysis

Nitisha Jain

Hasso-Plattner-Institut, University of Potsdam, Germany
nitisha.jain@hpi.de

Abstract. Knowledge graphs are widely used for systematic representation of real-world data. Large-scale, general purpose knowledge graphs, having millions of facts, have been constructed through automated techniques from publicly available datasets such as Wikipedia. However, these knowledge graphs are typically incomplete and often fail to correctly capture the semantics of the data. This holds true particularly for domain-specific data, where the generic techniques for automated knowledge graph creation often fail due to several challenges, such as lack of training data, semantic ambiguities and absence of representative ontologies. The focus of this thesis is on automated knowledge graph construction for the cultural heritage domain. The goal is to tackle the research challenges encountered during the creation of an ontology and a knowledge graph from digitized collections of cultural heritage data. This paper identifies the specific research problems for these tasks and presents a methodology and approach for a solution, along with preliminary results.

Keywords: knowledge graphs · ontology learning · cultural heritage.

1 Introduction

Knowledge graphs (KG) have become fairly common as structured, machine readable repositories of data. Several large KGs have been developed by industry and academia that are in widespread use for supporting downstream applications such as search and question answering. Knowledge graphs rely on an underlying schema or *ontology* that consists of the concepts (that define the *type* of the entities) and the possible relations between them. The ontology holds the key to the semantic meaning of the facts in a KG and dictates the logical rules as well as restrictions for populating the KG. There have been several efforts at automatic construction of general purpose knowledge graphs by extracting information from the Web [2, 19]. However, the resulting KGs are rarely fully correct and never complete in their coverage [10]. This problem is further exacerbated in domain-specific use cases. General purpose KGs constructed from Web sources cover a wide range of domains and therefore they cannot be expected to be comprehensive and semantically aligned to any single domain in particular. In order for KGs to be useful for a specific domain, it is essential to

have a semantically-rich and comprehensive representation of the domain in the KG. For instance, the most important concepts and relations differ from one domain to the other, such as Bank and Loans for the financial domain and names of Proteins and Genes. for the biomedical domain. Due to this, the underlying ontologies in general KGs are insufficient for semantic representation of specific domains. Gold standard annotated datasets required for the training as well as evaluation of automated techniques are also largely absent for domain-specific tasks. As a result, knowledge graphs end up having specific tailor-made construction pipelines for different domains, while the domain ontologies are largely manually designed with the help of expensive human expertise.

In order to motivate and explore these research problems, we consider cultural heritage as a representative domain. We are working in collaboration with the Wildenstein Plattner Institute[1] that was founded to promote scholarly research on cultural heritage collection. A wealth of information is buried in large collections of recently digitized art resources. In these resources, cultural objects such as artworks, auctions, art collections, artistic movements etc. are often mentioned within semi-structured or unstructured texts. Identification of the mentions of these cultural objects as named entities and establishing their relations can facilitate search and browsing in digital resources, help art historians to track the provenance of artworks and enable wider semantic text exploration for digital cultural resources. However, extraction of this information to construct a representative *art* knowledge graph is a non-trivial task.

In this thesis, we identify the challenges of designing a framework for constructing a domain-specific KG in an automated manner. In particular, we examine how to automate the process of design of an ontology for a new domain as well as populate a KG based on this ontology with domain relevant facts via automated techniques. We explore the role of modern deep learning techniques for the different tasks of KG construction, including named entity recognition (NER), linking and relation extraction. Our goal is to devise methods and techniques for knowledge representation that perform well for the cultural heritage domain, while also being sufficiently robust and generic to be applicable to other domains.

1.1 Domain-Specific Challenges

Cultural heritage data is vastly heterogeneous and comprises of multiple topics, multiple languages as well numerous different text formats ranging from structured tabular data to long passages of unstructured text descriptions. Data obtained from historical archives also poses significant linguistic challenges in terms of outdated vocabularies and phrases, such that the modern natural language processing tools are unable to perform well for these texts [9]. In the absence of large annotation datasets, the adaptation of existing solutions faces many challenges, some of them being unique to this domain. As an example, if we consider the task of NER, existing state-of-the-art tools fail to recognize

[1] https://wpi.art/

important entities of the cultural heritage domain, such as artworks. Due to the ambiguities that are inherent in artwork titles, the identification of their mentions from texts is a challenging task and requires significant domain expertise to tackle. Consider the painting with the title '*Head of a woman*' — such phrases can be hard to get distinguished as named entities from the surrounding text due to their generality. Even the presence of typical formatting cues such as capitalization, quotes, italics or boldface fonts cannot be assumed or guaranteed, especially in digitized texts obtained from scans of art historical archives. The issue of noisy data due to OCR limitations further exacerbates the challenges for automated text analysis for the cultural heritage domain [18].

Fig. 1. Dataset Samples

1.2 Dataset

A large collection of digitized art historical documents has been provided by our project partners as a representative cultural heritage dataset. The dataset consists of texts in many different languages including English, French, German, Italian, Dutch, Spanish, Swedish and Danish among others. The collection comprises of different types of documents: auction catalogues, art books related to particular artists or art genres, catalogues of art exhibitions and other documents. The auction and exhibition catalogues contain semi-structured and unstructured texts that describe artworks on display, mainly paintings and sculptures. Art books may contain more unstructured text about the origins of artworks and their creators. For reference, a few sample documents from a similar collection of digitized exhibition catalogues[2] and historical art journals[3] are shown in Fig. 1.

[2] https://digi.ub.uni-heidelberg.de/diglit/koepplin1974bd1/0084,0095
[3] https://digi.ub.uni-heidelberg.de/diglit/studio1894/0019

2 Related Work

This work builds upon research in several domains. The relevant previous works and their limitations are briefly discussed here.

Semantic web and cultural heritage. With the principles of linked open data[4] gaining momentum in the cultural heritage domain [21], there has been a recent surge in the availability of digitized cultural data.Initiatives such as OpenGLAM[5] and flagship digital library projects such as Europeana[6] and Digital Public Library of America[7] aim to enrich open knowledge graphs with cultural heritage data by improving the coverage of the topics related to the cultural domain. Several efforts have been made to digitize historical archives and collections [8]. This is especially true for the art domain where a large collection of raw texts are yet to be explored that could benefit greatly from a systematic representation of the information in the form of a KG. Although there is previous work on creating ontologies and knowledge repositories for several specific use cases [6, 12], yet a comprehensive method for automatically constructing an *art* knowledge graph has thus far eluded this domain. This thesis aims to identify and overcome the unique challenges of the cultural heritage domain in order to automate the ontology and KG creation.

Ontology learning. Ontology learning or ontology inference has been a subject of active research. Towards this goal, previous works have focused on automatic taxonomy induction from structured and unstructured texts [23]. However, these approaches suffer from low coverage and do not scale well to domains with noisy datasets, thus requiring manual cleanup efforts. A number of tools for building ontologies from large datasets have also been developed [4, 17]. Inspite of existing frameworks, ontology construction for specialized domains still requires extensive collaboration between ontology engineers and domain experts for enabling accurate reasoning and knowledge inferencing. Automated methods for inferencing and extending domain ontologies is one of the research questions that will be addressed in this thesis.

Knowledge graph construction. Due to the popularity of knowledge graphs, automated KG construction has garnered a lot of attention from the research community. Large multi-lingual KGs such as Yago [16] and DBpedia [14] have been generated by leveraging Wikipedia for data and schema derivation. There have been several efforts towards automatic construction of general purpose KGs from the Web based on machine learning techniques [2, 19]. However, automated KG construction techniques suffer from a number of shortcomings in terms of their coverage and scalability. Generic techniques fail to achieve comparable performance for domain-specific datasets, particularly for cultural heritage collections. Though there have been some efforts in this direction [3], previous work on automated KG construction for the cultural heritage domain is relatively sparse and therefore, the main focus of this thesis.

[4] Linked Open Data: `http://www.w3.org/DesignIssues/LinkedData`
[5] OpenGLAM: `http://openglam.org`
[6] Europeana: `http://europeana.eu`
[7] DPLA: `https://dp.la/`

3 Problem Statement

The main goal of this thesis is to *enable automated construction of a domain-specific, semantically-rich knowledge graph from cultural heritage datasets.*

The construction of a knowledge graph from any data source involves ontology design as well as several information extraction tasks including named entity recognition (NER), entity linking and relation extraction. While these tasks are already subject to active research, the construction of an *art* knowledge graph faces domain-specific challenges and needs customized solutions for several research problems. We identify the following research questions as the focus of this thesis:

How can a domain-specific ontology be learnt automatically from data? Ontology design and construction is one of the first and most important steps for KG construction, yet it has largely remained a manual task, particularly for new domains. In order to create a KG from domain-specific data, experts are sought out to manually build a suitable representative ontology. General purpose ontologies, such as those that are included in DBpedia and Yago, may already contain a few concepts that are relevant for the domain and thus could be borrowed. However, to encompass all aspects of the domain, especially with reference to a specific dataset, the extension of the existing ontologies becomes essential. There are several ontologies that are have been designed for the semantic representation of specific cultural heritage datasets. For example, the OpenART ontology [1] describes a research dataset about London's art world. The CIDOC-CRM [5] is a well-known ontology that provides a description of heterogeneous cultural heritage information. However, these ontologies have been largely designed and derived from underlying datasets via manual efforts that could be laborious and expensive. Our problem statement is to enable automated ontology learning with the help of domain-specific datasets and existing ontologies in the context of the cultural heritage domain.

How can we extract artwork titles from cultural heritage data collections through named entity recognition? Titles of artworks, such as paintings and sculptures, are one of the most important entities in cultural heritage. It is common to have generic artwork titles such as '*Girl before a mirror*' (by Pablo Picasso) and abstract titles such as *'untitled'*, making it hard to identify such titles as named entities. Most existing NER efforts are restricted to only a few common categories of named entities, i.e., *person*, *organization*, *location*, and *date*. Fine-grained NER or FiNER aims to classify the entities into several more entity types [15] which is essential for domain-specific NER. However, previous works on FiNER are not specifically catered to the cultural heritage domain and therefore, do not explicitly identify artwork titles as a named entity type. For the construction of an *art* KG, we want to identify the mentions of artworks, as important named entities, from cultural heritage collections.

How can cultural heritage entities be connected by meaningful relations? Understanding the relations between the various entities of a domain is essential for semantic analysis. A comprehensive *art* knowledge graph will not only consist of artwork and artist entities, but also include cultural institutions

(such as museums and galleries), art styles and movements, auction and exhibition events related entities such as auction houses, exhibition venues, artwork owners etc., along with specific attributes and relations. A domain ontology can act as a guide for automated relation discovery by restricting the possible types of relations between two entities. E.g., an artwork can be connected with a museum through *exhibited* or *acquired* relation but not with *created* relation. Existing KGs (such as Wikidata that contains almost 15,000 artwork entities) can be leveraged to obtain an initial set of relations to train machine learning models for further inference. However, due to their skewed and incomplete representation of the domain(limited to instances of only a few popular entities such as artworks and artists), an accurate and comprehensive representation of domain-specific datasets is not possible by merely re-using the existing KGs, but requires the construction of a domain-specific KG. In order to make such a KG useful for semantic exploration by the domain experts, further enrichment is desirable, which leads to the next research question.

How can enrichment of an art knowledge graph enable efficient semantic exploration? The augmentation of the cultural heritage entities and relations with additional attributes can prove useful for exploration by art experts. Artwork entities can be enriched with provenance information to facilitate the tracking of their history and origins, whereas relations between artists can be enhanced with data about their influence on each other's work. Taking advantage of the multilingual texts present in cultural heritage collections, the KG can be enriched with multilingual labels for different entities, especially artworks. Further, clustering techniques can be used for inference tasks such as identification of art styles for artists, this insight can be added back to the KG for discovery and analysis by art historians. We focus on the enrichment and refinement of the KG to alleviate its usefulness for semantic exploration of cultural heritage.

4 Research Methodology and Approach

The overall methodology adopted for addressing the research problems in this thesis consists of the following steps:

1. Identification and exploration of the research space, including work on the same task from other domains.
2. Investigation of the limitations of any existing solutions and formulation of the challenges specific to cultural heritage domain.
3. Formalization and implementation of the possible solutions.
4. Evaluation of performance as compared to state-of-the-art techniques.
5. Iterative improvement in performance and usability based on feedback from our project collaborators as well as art historians.

We discuss more on the planned approach for the ongoing research tasks in this section, i.e. automated ontology learning and NER for artwork titles.

Automated ontology learning for domain-specific datasets. In spite of several previous efforts, automated ontology construction is still considered to

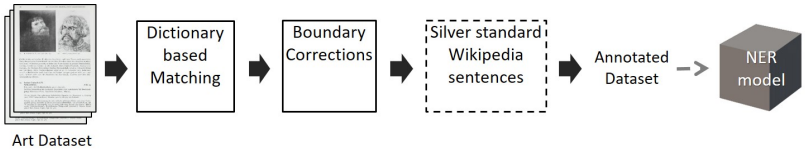

Fig. 2. First approach for NER for Artwork titles

be an open problem. We propose to build an ontology for cultural heritage domain with the help of *knowledge graph embeddings*. KG embedding models based on general purpose KGs, such as Yago and DBpedia, have gained significant attention in the past decade [22]. They have been shown to successfully improve KGs by performing link prediction, entity typing and resolution. However, most KG embeddings only model entity triples and ignore the rich semantic information which comes from the ontological triples that are already present in many modern KGs. The addition of ontological information to KG embedding models can improve their performance for KG completion as well as extend their utility towards completion of the underlying ontological structure [7, 11]. We propose that general KG embeddings enriched with domain-specific ontological information can be used for predicting incomplete ontological triples, thereby helping in ontology learning. We envision to leverage the existing ontological information pertaining to cultural heritage that is present in frameworks such as CIDOC-CRM (or even Yago and DBpedia) and extend these ontologies to comprehensively describe previously unseen datasets.

Named entity recognition for artwork titles. Towards building an *art* KG, we have started with the task of identifying mentions of artworks as named entities from digitized art archives. Recognizing the lack of annotated training datasets as one of the major bottlenecks, we are designing a framework for generating a large annotated corpus for training an NER model (Fig. 2). Firstly, existing art resources, that are integrated in popular knowledge bases, such as Wikidata[8] were leveraged to create a large entity dictionary or *gazetteer* of around 15,000 artwork titles. By matching the titles in the entity dictionary with the text, we obtained precise annotations of named entity type *artwork* from the underlying dataset, from which the NER model was able to learn useful features. Annotation errors due to partial matching of named entities were handled by heuristics-based boundary corrections to obtain higher recall of annotations. Further, to enable an NER model to learn from the textual patterns present in the dataset for identification of artworks, we plan to further augment the training dataset with clean and well-structured silver standard annotations that can be derived from Wikipedia articles [20]. Through this process, we aim to generate a large corpus of annotated data via automated approaches from any art corpus for retraining existing NER tools and identify mentions of artwork titles from art collections.

[8] `https://www.wikidata.org`

5 Evaluation Plan

The evaluation of the quality of a knowledge repository geared towards the cultural heritage domain must be determined by its usefulness for the domain experts. As such, we hope to enlist the assistance of our project partners to provide the necessary feedback as well as critical comments at different stages of knowledge graph construction. This is particularly important for ontology creation, where the quality of an inferred ontology can best be judged by domain experts. In order to perform an empirical evaluation of our proposed method to learn ontologies using knowledge graph embeddings, gold standard test data can be created by deleting some concepts from an existing ontology and inferring these concepts to test the effectiveness of the method.

For judging the quality of the overall KG from an information extraction point of view, we plan to consider two aspects - *completeness* in terms of the coverage of the facts, and *correctness* in terms of number of erroneous facts. The precision, recall and F1 scores will be used for quantifying the completeness whereas the accuracy measure can be used for correctness. Although a deterministic measurement of the completeness of the KG is difficult due to the open world assumption, we plan to make estimations based on the coverage of facts that can be extracted from pre-identified texts (such as Wikipedia articles). For performing these intrinsic evaluations, we plan to manually create a gold standard dataset against which the scores will be calculated. We also plan to perform the extrinsic evaluation for the KG in the context of domain-specific use cases to determine whether the KG can support certain desirable downstream tasks such as search and retrieval of artwork entities and other semantic tasks including named entity disambiguation, semantic similarity and pattern mining.

6 Preliminary Results

In this section, we present the first results from our research efforts on NER for artworks that have been peer-reviewed [13]. As discussed in Section 1.2, the dataset for this work consisted of a sizeable collection of digitized art historical documents. After initial pre-processing the dataset consisted of 19,310,429 sentences, which was then transformed into annotated NER data with the first two steps of the approach as described in Section 4 (Fig.2). The number of annotations and unique entities in the training dataset were respectively - 413,932 and 24,966. In order to measure the impact of the quality of the training data on NER performance, we trained the baseline NER model for the new entity type *artwork* on the annotated training dataset and evaluated the trained model with the help of a manually created test dataset. The performance of the re-trained NER model in terms of precision, recall and F1 scores was evaluated with strict[9] as well as relaxed[10] metrics (based on exact and partial boundary matches).

[9] `https://www.clips.uantwerpen.be/conll2003/ner/`

[10] `https://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html`

Table 1. Performance of NER Models Trained on Different Datasets

	Strict				Relaxed			
NER model	P	R	F1	Acc	P	R	F1	Acc
Baseline	.14	.06	.08	.24	.22	.08	.12	.37
Re-trained	.23	.22	.23	.61	.39	.41	.40	.68

The preliminary results as shown in Table 1 have demonstrated notable improvement in performance for the NER models that were trained with annotated data as compared to the baseline performance. Though the improvements are encouraging, the absolute numbers are still low for the NER model to be useful in practice. Thus, we are exploring further improvement in performance by the addition of contextual features to the training dataset, such as annotations for artist names and art styles.

7 Conclusion

The main goal of this thesis is the automated construction of domain-specific knowledge graphs and ontologies. To this end, we consider the cultural heritage domain and adapt information retrieval tasks to overcome specific domain challenges. So far, we have studied related work and defined a methodology that will guide our research efforts throughout this work. We have partially addressed the problem of named entity recognition for artworks in a cultural heritage collection and obtained promising preliminary results that highlight the potential for the practical applications of this work.

Acknowledgement I am thankful to my advisor Ralf Krestel for his feedback and Felix Naumann and Fabian Suchanek for their valuable comments. I would also like to thank Elena Demidova for guidance and suggestions during revisions.

References

1. Allinson, J.: Openart: Open Metadata for Art Research at the Tate. Bulletin of the American Society for Information Science and Technology **38**(3), 43–48 (2012)
2. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E.R., Mitchell, T.M.: Toward an Architecture for Never-Ending Language Learning. In: Proc. of the 24th AAAI Conference on Artificial Intelligence. pp. 1306–1313 (2010)
3. 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: Proc. of the International Semantic Web Conference. pp. 36–52. Springer (2019)
4. Cimiano, P., Völker, J.: Text2Onto. In: Proc. of the International Conference on Application of Natural Language to Information Systems. pp. 227–238 (2005)
5. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M.: Definition of the CIDOC Conceptual Reference Model. ICOM/CIDOC Documentation Standards Group. CIDOC CRM Special Interest Group **5** (2008)

6. van Dalen-Oskam, K., de Does, J., Marx, M., Sijaranamual, I., Depuydt, K., Verheij, B., Geirnaert, V.: Named Entity Recognition and Resolution for Literary Studies. Computational Linguistics in the Netherlands Journal **4**, 121–136 (2014)
7. Diaz, G.I., Fokoue, A., Sadoghi, M.: EmbedS: Scalable, Ontology-aware Graph Embeddings. In: Proc. of the EDBT Conference. pp. 433–436 (2018)
8. Dijkshoorn, C., Jongma, L., Aroyo, L., Van Ossenbruggen, J., Schreiber, G., ter Weele, W., Wielemaker, J.: The Rijksmuseum Collection as Linked Data. Semantic Web **9**(2), 221–230 (2018)
9. Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F.: Diachronic Evaluation of NER Systems on Old Newspapers. In: Proc. of the 13th Conference on Natural Language Processing (KONVENS 2016). pp. 97–107 (2016)
10. Galárraga, L., Razniewski, S., Amarilli, A., Suchanek, F.M.: Predicting Completeness in Knowledge Bases. In: Proc. of the 10th ACM International Conference on Web Search and Data Mining. pp. 375–383 (2017)
11. Hao, J., Chen, M., Yu, W., Sun, Y., Wang, W.: Universal Representation Learning of Knowledge Bases by Jointly Embedding Instances and Ontological Concepts. In: Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1709–1719 (2019)
12. Hellmund, T., Hertweck, P., Hilbring, D., Mossgraber, J., Alexandrakis, G., Pouli, P., Siatou, A., Padeletti, G.: Introducing the HERACLES Ontology—Semantics for Cultural Heritage Management. Heritage **1**(2), 377–391 (2018)
13. Jain, N., Krestel, R.: Who is Mona L.? Identifying Mentions of Artworks in Historical Archives. In: Proc. of the International Conference on Theory and Practice of Digital Libraries. pp. 115–122 (2019)
14. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web **6**(2), 167–195 (2015)
15. Ling, X., Weld, D.S.: Fine-Grained Entity Recognition. In: Proc. of the 26th AAAI Conference on Artificial Intelligence. pp. 94–100 (2012)
16. Mahdisoltani, F., Biega, J., Suchanek, F.: YAGO3: A Knowledge Base from Multilingual Wikipedias. In: 7th Biennial Conference on Innovative Data Systems Research. CIDR Conference (2014)
17. Navigli, R., Velardi, P.: Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. Computational Linguistics **30**(2), 151–179 (2004)
18. Rodriquez, K.J., Bryant, M., Blanke, T., Luszczynska, M.: Comparison of Named Entity Recognition Tools for Raw OCR Text. In: Konvens. pp. 410–414 (2012)
19. Shin, J., Wu, S., Wang, F., De Sa, C., Zhang, C., Ré, C.: Incremental Knowledge Base Construction using Deepdive. In: Proc. of the VLDB Endowment International Conference on Very Large Data Bases. vol. 8, p. 1310 (2015)
20. Tsai, C.T., Mayhew, S., Roth, D.: Cross-lingual Named Entity Recognition via Wikification. In: Proc. of the 20th SIGNLL Conference on Computational Natural Language Learning. pp. 219–228 (2016)
21. Van Hooland, S., Verborgh, R.: Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata. Facet publishing (2014)
22. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge Graph Embedding: A Survey of Approaches and Applications. IEEE Transactions on Knowledge and Data Engineering **29**(12), 2724–2743 (2017)
23. Yang, H., Callan, J.: A Metric-based Framework for Automatic Taxonomy Induction. In: Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 271–279. Association for Computational Linguistics (2009)