Generating Domain-Specific Knowledge Graphs: Challenges with Open Information Extraction

Nitisha Jain[1][0000−0002−7429−7949], Alejandro Sierra-Múnera[1][0000−0003−3637−4904], Maria Lomaeva[2], Julius Streit[1], Simon Thormeyer[1], Philipp Schmidt[1], and Ralf Krestel[3,4][0000−0002−5036−8589]

[1] HPI - Hasso Plattner Intitute, Potsdam, Germany
[2] Potsdam University, Potsdam, Germany
[3] ZBW - Leibniz Centre for Economics, Kiel, Germany
[4] Kiel University, Kiel, Germany

Abstract. Knowledge Graphs (KGs) are a popular way to structure and represent knowledge in a machine-readable way. While KGs serve as the foundation for many applications, the automatic construction of these KGs from texts is a challenging task where Open Information Extraction techniques are prominently leveraged. In this paper, we focus on generating a domain-specific knowledge graph based on art-historic texts from a digitized text collection. We describe the combined use and adaption of existing open information extraction methods to build an art-historic KG that can facilitate data exploration for domain experts. We discuss the challenges that were faced at each step and present detailed error analysis to identify the limitations of existing methods when working with domain-specific corpora.

Keywords: knowledge graphs · open information extraction · domain-specific texts

1 Introduction

Knowledge Graphs (KGs) have gained considerable popularity in both academia and industry. They are employed to represent information in a structured format after extraction from large collections of heterogeneous, diverse, and unstructured documents [19]. These KGs can then be used for downstream tasks, such as question answering, logical inference, recommendation, or information retrieval. Besides general KGs that aim to capture generic knowledge about real-world data, such as DBpedia [27] and Wikidata [39], domain-specific KGs have become important for targeted domains [25]. They have been leveraged to support multiple information-based applications, e.g., in the context of health and life sciences [13], news search [34] or fact checking [8].

There have been several efforts towards automatic construction of general purpose knowledge graphs from the Web based on machine learning techniques [35,5]. In the absence of a pre-specified list of relations for performing pattern-based extractions, Open Information Extraction (Open IE) is a popular ap-

proach, where a large set of relational triples can be extracted from text without any human input or domain expertise [15]. Several Open IE techniques have been proposed to build and populate knowledge graphs from free-form texts [16,43,2,10,18,26]. However, these methods for automated knowledge base construction suffer from a number of shortcomings in terms of their coverage [17] and applicability to specific domains [25]. Existing techniques that exhibit state-of-the-art results on standard, clean datasets fail to achieve comparable performance for domain-specific datasets, e.g., in the art-historic domain where the data often consists of highly heterogeneous and noisy collections [22].

KG for Art. The art and cultural heritage domain provides a plethora of opportunities for knowledge graph applications. An art knowledge graph can enable art historians, as well as interested users, to explore interesting information that is hidden in large volumes of text in a structured manner. With a large variety of diverse information sources and manifold application scenarios, the (automated) construction of task-specific and domain-specific knowledge graphs becomes even more crucial for this domain. In contrast to general purpose KGs, a KG for the art domain could comprise a specific set of entity types, such as artworks, galleries, as well as relevant relations, such as influenced_by, part_of_movement etc., depending on the specific task and on the specific text collection. The important entities and relations might also differ across different document types, such as auction catalogues, exhibition catalogues, or art magazines. On one hand, a general purpose, art-oriented ontology may not be well-suited and comprehensive enough for specific data collections. On the other hand, designing a custom ontology for the different art corpora would be a challenging and expensive task due to the need for significant domain expertise. In the past, several attempts have been made at creating KGs for art and related domains [41,21,6], with the most recent one by Castellano et al. [7]. However, a systematic method for the construction of a knowledge graph based on a collection of art-related documents without a well-defined ontology has not been proposed thus far.

Goals. In this paper, we describe an ongoing project[5] for the automatic construction of a knowledge graph based on a large, private archive of art-historic documents. Instead of relying on existing ontologies to dictate the information extraction process (that might restrict the scope of the entities and relations that could be extracted from the text when the ontology is not hand-crafted for the specific dataset) we decided to pursue the schema-less Open IE approach in this work. We present the results from our exploration of existing Open IE techniques to generate structured information and discuss our insights in terms of their shortcomings and limited applicability when deployed for noisy, digitized data in the art domain.

We make the following contributions in this paper: (i) Construct a domain-specific knowledge graph based on a collection of digitized art-historic documents.

[5] https://hpi.de/naumann/projects/web-science/ai4art.html

(ii) Describe the process of automated construction of the KG with Open IE techniques. (iii) Analyze and discuss the challenges and limitations for the adaptation of Open IE tools to domain-specific datasets.

2 Related Work

With the availability of digitized cultural data, several previous works have proposed KGs for art-related datasets [41,21,6,31]. Arco [6] is a large Italian cultural heritage graph with a pre-defined ontology that was developed in a collaborative fashion with contributions from domain experts all over the country. While the Arco KG is quite broad in its coverage, Ardo [40] pertains to a very specific use case of multimedia archival records. Similarly, the Linked Stage Graph [36] was developed as a KG specifically for storing historical data about the Stuttgart State Theater. Increasingly, the principles of linked open data[6] have also been widely adopted within the cultural heritage domain for facilitating researchers, practitioners and generic users to study and consume cultural objects. Notable examples include the CIDOC-CRM [30], the Rijksmuseum collection [12], the Zeri Photo Archive[7], OpenGLAM [37] among many others. Most related to our work is the ArtGraph [7] where the authors have integrated the art resources from DBpedia and WikiArt and constructed a KG with a well-defined schema that is centered around artworks and artists. While all these works are concerned with KGs and ontologies for specific art-related corpora, they have leveraged a schema for representing the information and are not concerned with the challenges of a schema-free extraction process, which is the main focus of this work.

Open IE approaches extract triples directly from text, without an explicit ontology or schema behind the extraction process. Several works have been proposed in the past. TextRunner [43] relies on a self supervised classifier which determines trustworthy relationships with pairs of entities, while Reverb [16] uses syntactical and lexical constraints to overcome incoherent and uninformative relationships. ClausIE [10] relies heavily on dependency parsing to construct clauses from which the propositions will be extracted. In this work, we have leveraged the Stanford CoreNLP OpenIE implementation [28,2] that uses dependency parsing to minimize the phrases of the resulting clauses, and was originally evaluated in a slot filling task.

The construction of domain-specific KGs has been the subject of investigation in previous works for various domains, e.g. software engineering [45], academic literatures [20], and more prominently, the biomedical domain [44,3,14]. However, the previously proposed automated methods are not directly applicable for the arts and cultural heritage domain, where unique challenges with respect to the heterogeneity and quality of data are prevalent. This work identifies and discusses the particular difficulties encountered while applying existing information extraction techniques to art-related corpora.

[6] Linked Open Data: http://www.w3.org/DesignIssues/LinkedData
[7] https://fondazionezeri.unibo.it/en

3 Automated Construction of Art-historic KG

In this section, we describe our underlying art-historic dataset as well as the steps employed for the automated extraction of information (in form of triples) to construct an art-historic knowledge graph. Fig. 1 shows an overview of this process.

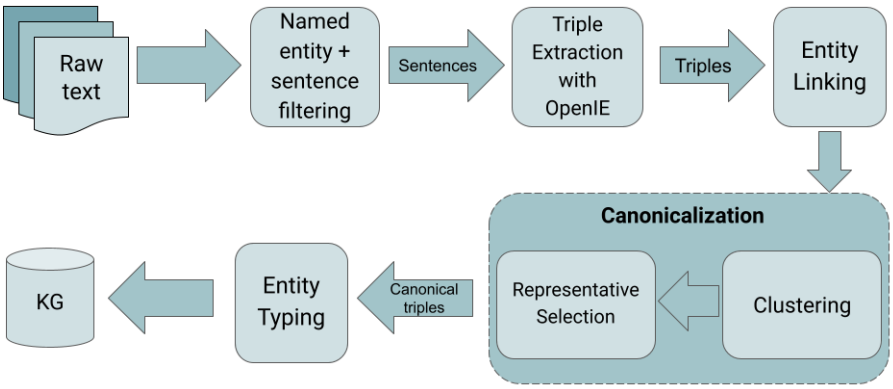


Fig. 1: Construction of art-historic KG

3.1 Dataset

For this work, we are working with a large collection of recently digitized art-historical texts provided by our project partners. This collection consists of a variety of heterogeneous documents including auction catalogs, exhibition catalogs, art books, etc. that contain semi-structured as well as unstructured texts describing artists, artworks, exhibitions and so on. Art historians regularly study these data collections for art-historical analysis. Therefore, a systematic representation of this data in the form of a KG would be a valuable resource for them to explore this data swiftly and efficiently. The whole collection is quite large ($\approx$ 1TB of data), in order to restrict the size of the dataset for a proof-of-concept of our KG construction process, a subset of this dataset pertaining to information about the artist *Picasso* was chosen. The decision of choosing an artist-oriented subset of the collection enabled us to better understand the context and evaluate the triples that were obtained throughout the process of KG construction. The data was filtered by querying the document collection using the keyword query *'Picasso'*, resulting in 224,469 entries (where each entry corresponds to a page of

the original digitized corpus) containing the term *'Picasso'*. Due to the filtering, each entry is an independent document, in the sense that the neighboring entries do not always represent the correct context. This led to some of the entries in our dataset containing incomplete sentences at the beginning or the end of a page. One such example is an entry starting with *'to say47—Picasso never belittled his work, until . . . '* where the tokens *'to say'* belong to a sentence which started in a different entry, that might no longer be a part of the dataset under consideration. It is important to note that in the same example we can see more noise, e.g., numbers are mixed in between words in the digitized version of the text. This noise in the dataset was introduced by the optical character recognition (OCR) process during the digitization of the documents (performed in a prior step by the data providers). In general, the dataset contains full sentences, such as *'Matisse's return to the study of ancient and Renaissance sculpture is significant in itself.'*, as well as short description phrases, figure captions or footnotes such as *'G. Bloch, Pablo Picasso, Bern, 1972, vol. III, p.142'*.

3.2 Finding Named Entities

As a first step, it was interesting to inspect if the named entities present in the corpus could be easily identified. A dictionary-based approach to find the named entities would identify the mentions with a high precision, but at the cost of very low recall by ignoring many potentially interesting entities to be discovered in the corpus. Therefore, we chose to follow a machine learning approach to named entity recognition (NER). Generic NER tools work very well for the common entity types, such as person, location, organization and so on, though fine-grained or domain-specific entities are harder to identify [23]. We employed the SpaCy library[8] for finding named entities since its pre-trained models includes a Work_Of_Art category that could potentially identify the entities that are important in the art domain (this could encompass mentions of paintings, books, statues etc.). Excluding the cardinal entities in order to reduce noise, the SpaCy library with the pre-trained *'en_core_web_trf'* model was used to identify the following entity types - Work_Of_Art, Person, Product, ORG, LOC, GPE and NORP, which showed reasonably good results. The process of NER enabled us to filter out any sentences without any entity mention since such sentences were likely to have no useful information for the KG construction. Thus, the NER step helped with pruning the dataset for further processing, as well as improving the quality of the resulting KG.

3.3 Triple Extraction

After obtaining informative sentences from the previous step, we employed Open IE tools to extract the triples from them. It is important to note that while there are some art-related ontologies proposed in previous works such as Arco [6] and ArDo [40], none of them are suitable for our corpus since they are very specific

[8] https://spacy.io/usage/v3

to the datasets they were designed for. Other general ontologies such as CIDOC-CRM are, on the other hand, too broad and would not be able to extract novel and interesting facts from a custom and heterogeneous corpus such as ours, where the entities and relations among them are not known before hand. In the absence of such an ontology specifically designed for the description of art-historic catalogs, we choose to employ open information extraction techniques for the construction of our KG in order to broaden the scope and utility of the extracted information.

To this end, we ran the Stanford CoreNLP OpenIE annotator [28,33] to extract ⟨subject, predicate, object⟩ triples from the sentences. A total of 5,057,488 triples were extracted in this process, where multiple triples could be extracted from a single sentence. Another round of filtering was performed at this stage, where any triples that did not contain a named entity in the subject or object phrase were removed. Additionally, duplicate entries and triples with serial numbers as entities were also ignored. Some examples of triples that were removed are: ⟨*we, have, good relationship*⟩, ⟨*i, be, director*⟩, ⟨*brothel, be in, evening*⟩, ⟨*drawings, acquired, work*⟩. A total of 160,000 triples remained, a valid triple at this stage looked like ⟨*P. Picasso*, *is*, *artiste*⟩.

3.4 Entity Linking

Once the triples were extracted, the entity linking component of the Stanford CoreNLP pipeline [28] was used to link the entities. This component uses Wiki-Dict as a resource, and uses the dictionary to match the entity mention text to a specific entity in Wikipedia. Since the entities in our dataset were present in multiple different surface forms, this step allowed us to partially normalize the entities and identify the unique entities. Though the number of entities was reduced as a result, the total number of triples remained the same. Note that this linking could only map entities to their Wikipedia counterpart if the entity was found as a subject or object in a triple. In many cases though, the subject and object were noun phrases instead of obvious entities, for which this kind of linking did not really work. This process was still quite useful as around 108,841 out of 337,100 entities were successfully linked to their Wikipedia form (leading to 8,369 unique entities). Some of the most frequent entities found in the dataset (along with their frequencies) were: (*Pablo_Picasso*, 11219), (*Paris*, 2178), (*Artist*, 1904), (*Henri_Matisse*, 1769), (*Georges_Braque*, 1352).

3.5 Canonicalization

One of the main challenges when constructing a KG through Open IE techniques, is that of canonicalization. Multiple surface forms of the same entity or relation might be observed in the triples extracted with Open IE techniques in the form of noun phrases or verb phrases that need to be identified and tagged to a single semantic entity or relation in the KG. Since the triples extracted from our dataset via Open IE method comprised many noisy phrases, as well as new entities, such as titles of artworks, that may not be available for mapping

in existing databases, entity linking techniques would not suffice in this case. Different from entity linking (that can only link entities already present in external KGs), canonicalization is able to perform clustering for the entities and relations that may not be present in existing KGs, by labelling them as OOV (out of vocabulary) instances. In this work, we chose to perform canonicalization with the help of CESI [38] which is a popular and openly available approach for this task. The CESI approach performs clustering over the non-canonicalized forms of noun phrases for entities and verb phrases for the relations. It leverages different sources of side information for noun phrases and relation phrases such as entity linking, word senses and rule-mining systems for learning embeddings for these phrases using the HolE [29] knowledge graph embedding technique. The clustering is then performed using hierarchical agglomerative clustering (HAC) based on the cosine similarity of the phrase embeddings in vector space. In this manner, different phrases for the same entity or relation were mapped to one canonicalized form for including in the KG. In total, we obtained 3,789 entity clusters and 3,778 relation clusters from the CESI approach that contained two or more terms.

Representative Selection. An important step in the CESI approach is the assignment of representatives for the clusters obtained for the noun and relation phrases. This is decided by calculating a weighted mean of all the cluster members' embeddings in terms of their frequency of occurrence. The phrase closest to this mean is selected as the representative. However, this technique did not work well for our domain-specific and noisy dataset and many undesirable errors were noticed. For example, an entity cluster obtained from CESI was: *Olga_Khokhlova*, *olga*, *khokhlova*, *picasso*. Since *Picasso* is the most frequent entity in the dataset, it was chosen as representative by CESI, but this is clearly wrong since *Picasso* and *Olga* are different entities. There were several other errors observed, e.g., all days of the week were clustered together in one cluster. This could be a result of the embedding and contexts of the days of the week to be quite similar, hence their vectors would end up together in the vector space. In other cases, the color *blue* occasionally showed up in a cluster of phrases related to color *red*, certain dates got clustered and certain related but not interchangeable words got clustered (*kill* vs *murder* vs *shot*). In some cases, the first name was being replaced by the incorrect full name (not every *david* is *david johnson*). To mitigate the above discussed errors, we had to perform manual vetting of the clusters for verification and selection of the correct cluster representatives which took around 2-3 person hours. During this process, certain clusters, where the entities were different, were removed (such as the cluster with days of the week). After this, the entities and relations were canonicalized as per their chosen cluster representatives leading to a total of 35,305 unique entities and 33,448 unique relations in the final KG[9].

[9] It is to be noted that existing canonicalization techniques such as CESI are largely optimized for canonicalization of entities and their performance is considerably worse for relations. We also observed similar results during our analysis.

3.6 Entity Typing

Since a schema or ontology was not employed to extract the triples from text, the entities in our KG do not have any entity types implicitly assigned to them. Therefore, we attempted to identify the types of as many entities in our graph as possible. With the help of NER, we assigned the types to the entities that were recognized in the triples. A total of 14,960 entities were typed with this technique to generic types such as Person, Product, ORG, LOC, GPE, NORP and Work_Of_Art, as well as numeric types such as Date, Time and Ordinal. Note that *Work_of_Art* is quite a broad category that includes artworks but also movies, books and various other art forms. Since artworks such as paintings and sculptures are one of the most important entities in our art-historic KG, it is worthwhile to identify the mention and type of these entities. However, generic NER process is neither equipped nor optimized to correctly identify such mentions. Thus, we additionally applied dictionary-based matching. This was done by compiling a large gazetteer of artwork titles by querying Wikidata with the help of the Wikidata Query Service[10] for the names of paintings and sculptures, retrieving approximately 15,000 artwork titles. In addition, we augmented our dictionary with the names of the *artwork* entities from the ArtGraph dataset [7] which contains more than 60,000 artworks derived from DBpedia and WikiArt. If a match was found for an entity in our KG in the compiled dictionary, the type was assigned as *artwork* accordingly. This led to the tagging of further 1,397 entities in our KG as artworks. The dictionary-based matching for artworks was particularly useful in the cases where it was able to correctly identify entities that were wrongly assigned as the *Person* type by NER, such as *la_donna_gravida*, *portrait_of_mary_cassatt* and *st._paul_in_prison*. Similar to artworks, we attempted to additionally identify the names of artists in our triples. While NER could only tag entities as *Person*, we used a dictionary of artist names from Wikidata to identify 656 unique artist entities in our data. These included names of artists such as *Piet Mondrian*, *Edvard Munch* and *Rembrandt*.

However, the process of entity typing described above is only able to identify and tag around half of the entities in our KG. Several domain and corpus-specific challenges acted as bottlenecks during this process. For example, even after filtering, some triples extracted from Open IE contained either subject or object noun phrases that were generic and did not correspond to any named entity. Examples of such phrases include *essay*, *anthology*, *periodical*, or *album* that are present in triples such as ⟨*album, be_shown_in, Paris*⟩. Without designing a custom ontology for this corpus, such entities cannot be hoped to be correctly typed.

The categorization of the relations in the KG is a particularly complicated task due to the wide variety of relations extracted from the Open IE process. Few of the most frequent relations in the KG are *will*, *be_in*, *have*, *show*, *paint*, *work* etc. We estimated that the types of the entities could be utilized to find patterns and link the most popular edges in the KG to the relations in existing

[10] https://query.wikidata.org/

Table 1: Statistics of the KG.

Attribute	Total Triples	Unique Entities	Unique Relations	Artworks	Artists
Count	147,510	35,305	33,448	1,397	656

graphs such as Wikidata or ArtGraph. However, preliminary analysis led to some interesting observations. Firstly, we noted the presence of multiple relations between pairs of entities in the KG. For example, *Picasso* and *June* are connected by various relations such as *will_be*, *work* and *take_trip_in* that were extracted from different contexts in the corpus and represent separate meaningful facts. Furthermore, in general, there are several different types of semantic relations between the popular entity types in our KG. For instance, two entities of the type *artist* are connected by several relations including *work*, *meet*, *know_well*, *be_with*, *friend_of* and *be_admirer_of*. While this variety indicates that a large number of interesting facts have been derived by Open IE in the absence of a fixed and limiting schema, normalizing the relations to improve the quality of the KG is a difficult task that is part of the ongoing and future work.

4 Art-historic Knowledge Graph

The statistics of the KG generated from the steps as described in the previous section are shown in Table 1.

4.1 Graph Features

After obtaining the refined set of triples for the first version of the art-historic KG, we performed a preliminary analysis of the graph to derive useful insights with the help of the NetworkX[11] package. To understand the graph structure, the number of disconnected components of the graph was measured before and after the canonicalization step. It was noticed that the number of disconnected components was reduced to around 1,500 (down from 2,500) after clustering with CESI. This indicates that canonicalization of entities and relations improved the quality of the knowledge graph by removing unnecessary disconnected parts that were created through redundant triples. Additionally, we also performed node centrality on the graph using eigenvector centrality [4] and link analysis using PageRank [32]. For both the measures, the node for *Pablo Picasso* was the most central. This confirms the property of the underlying dataset which is focused on *Picasso*. Other central nodes discovered were corresponding to popular words in the corpus such as *work*, *artist*, *painting* etc. Overall, it is promising to witness that centrality analysis of the generated KG conforms well regarding the main entities and topics of the underlying corpus. A hand-picked example of a subset of the neighborhood of the entity *Picasso* is shown in Fig. 2.

[11] https://pypi.org/project/networkx/

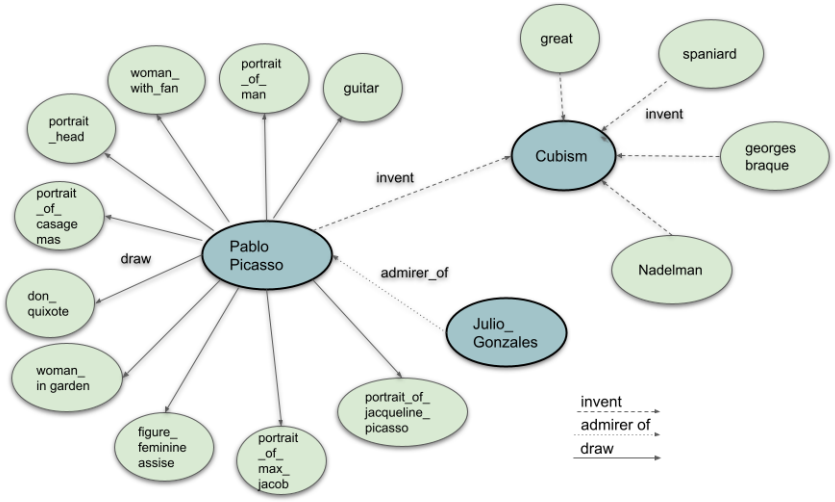


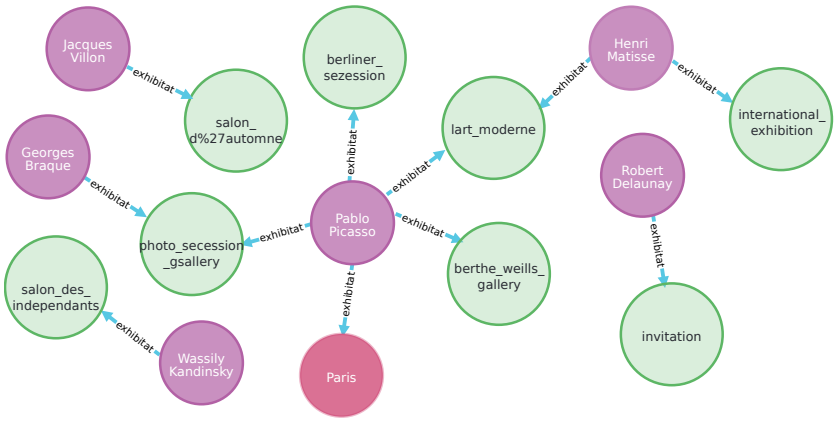
Fig. 2: Illustration of a subset of the KG

4.2 Evaluation

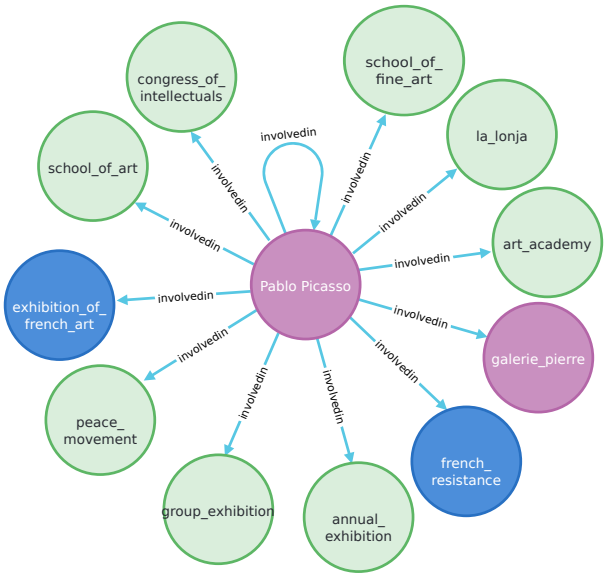
Due to the lack of any gold standard for direct comparison, the evaluation of the resulting KG proved challenging. While an absolute measure of the coverage of any KG is a non-trivial task due to the open world assumption [17], we attempted to perform limited evaluation in terms of the coverage of the KG in a semi-automated fashion. For this, we first created a subset of Wikidata [39] by querying for triples about the entity *Picasso* and used this as the knowledge graph for comparison. This is motivated by the fact that Wikidata contains high quality information about *Picasso* and the entity linking used in our pipeline performs the linking to Wikipedia (hence, Wikidata) entities. Therefore, it was likely to have a higher match between the surface forms of entities in our KG to the Wikipedia entities, as compared to other datasets such as DBpedia.

From the obtained Wikidata subset, 100 triples were randomly selected that related to information about *Picasso* as well as about museums that owned his works. Upon careful manual inspection (independently by three annotators) and resolution of conflicts with discussions, it was measured that the facts represented in 43% of these triples were also present in our KG as a direct match or in a different form with the same meaning. Notably, our KG was missing information about the museums that own Picasso's works, this is because our underlying corpus is also lacking comprehensive information on this topic. Therefore, triples relating to museums from Wikidata could not be matched. Additionally, we checked how many of our entities and entity pairs are written in exactly the same way as in the Wikidata graph. Overall, around 12% of entities and 10% of entity pairs in our graph have exact matches in Wikidata. These preliminary results are promising and point towards the need for a domain-oriented construction process

for further improvement of the art-historic KG. In particular, the precision of the triples in art-historic KG is more important to the users and therefore, factual verification for the triples that were extracted from our dataset but are not found in Wikidata needs to be conducted by enlisting the help of domain experts.



(a) Artists *exhibited at*. (corresponding query:
`MATCH p=(:Artist)-[r:exhibitat]->() RETURN p`)



(b) Picasso *involved in* various Art schools. (corresponding query:
`MATCH p=(s)-[r:involvedin]->() WHERE s.name="Pablo Picasso" RETURN p`)

Fig. 3: Examples of query results on the KG (node colours assigned by Neo4j).

4.3 Implementation

Taking cue from related work [7], we have encoded our KG data into Neo4j[12] which is a no-SQL graph database that provides an efficient way of capturing the diverse connections between the different entities of our knowledge graph. Additionally, the knowledge graph stored in the Neo4j database can be queried easily with the help of the Cypher language for enabling data exploration and knowledge discovery. Fig. 3 shows the results of a few example queries that can be executed on the KG - venues where *Picasso* and other artists had exhibited their work; and various art schools or movements where *Picasso* was involved. Further, Fig. 4 shows the persons and/or art styles that *Picasso* influenced or was influenced by. In some cases, interesting connections with other relevant entities are also retrieved, thus providing useful cues for further exploration of the data in the KG for domain experts as well as interested users.

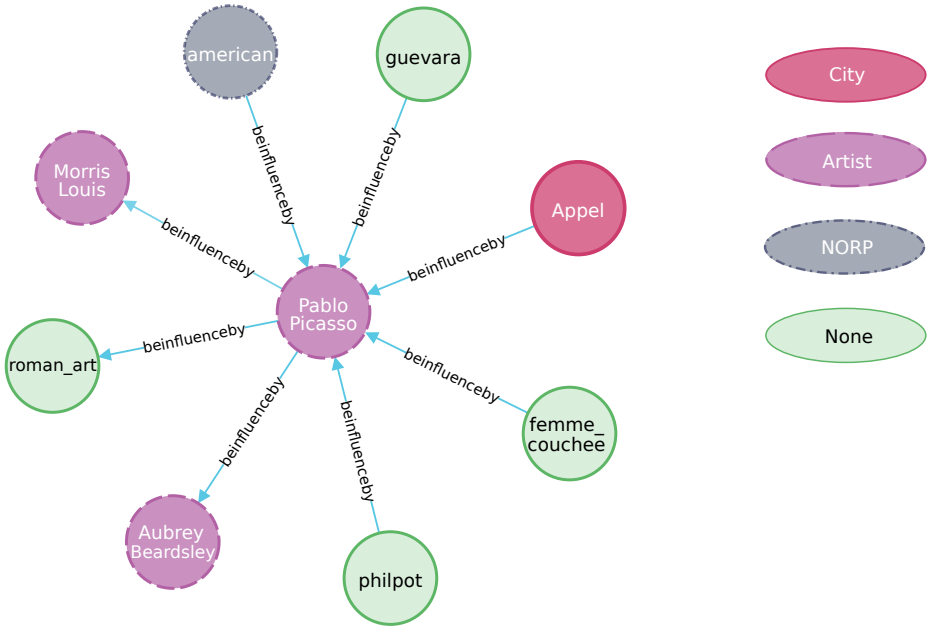


Fig. 4: Illustration of a subset of KG, depicting the influence of and on *Picasso* (corresponding query: `MATCH p=(s)-[:beinfluenceby]-(o) WHERE s.name="Pablo Picasso" RETURN p`)

[12] https://neo4j.com

Table 2: Examples of triples in the KG with their corresponding source texts

T1	⟨*The Third of May 1808, beIn, Madrid*⟩ At the center of the show, a room containing Francisco de Goya's **The Third of May 1808** *in* **Madrid** (1814), Édouard Manet's The Execution of Emperor Maximilian of Mexico (1868-69)...
T2	⟨*American, beInfluenceBy, Pablo Picasso*⟩ The more one examines Gorky's early works, the more they appear like Gorkys rather than like Picassos. Moreover, his unabashed borrowings can be seen as forward-looking: for an **American** to *be influenced by* **Picasso** in the heyday of American Scene painting was, art historian Meyer Schapiro points out, "an act of originality."
T3	⟨*Pablo Picasso, beInfluenceBy, Morris Louis*⟩ ...to Andrew Hudson, art critic of The Washington Post, for suggesting that **Pablo Picasso** has *been influenced by* **Morris Louis** and Kenneth Noland, two leaders of the "post-painterly" Washington, D.C.
T4	⟨*Guevara, beInfluenceBy, Pablo Picasso*⟩ It is probable that **Guevara** *was influenced by* **Picasso** to experiment with the encaustic technique, which had been practised in antiquity. Hot wax was used as a medium for mixing floral and vegetable dyes.
T5	⟨*Pablo Picasso, beInfluenceBy, Aubrey Beardsley*⟩ **Picasso** *was influenced* doubtless *by* **Aubrey Beardsley**, who had died in 1899 at the age of twenty-six, but then what an excellent influence it proved to be for this portrait !
T6	⟨*Appel, beInfluenceBy, Pablo Picasso*⟩ In artistic respect, one could also see, that Karel **Appel** *was* strongly *influenced* in this period, *by* **Picasso** and Miro.

5 Discussion and Error Analysis

Due to the source corpus being heterogeneous and noisy, the Open IE process led to a number of incorrect triples in the KG despite our best efforts to eliminate the noise at each step. Here, we perform a critical analysis and look deeper into the quality of the triples in the first version of the KG. For this, we sample few of the incorrectly extracted triples, to understand the nature of mistakes committed by the automated KG generation process. Table 2 presents some triples in the KG and the corresponding text snippets in the input data from which they were extracted.

In *T1*, even though the triple appears to be syntactically correct, the actual entity corresponds to the entire phrase *The Third of May 1808 in Madrid* which is an artwork, and thus the correct triples should relate this artwork to the corresponding artist *Francisco de Goya*, perhaps including the date *1814* as well. This example illustrates the difficulty of recognizing artwork titles, given that they usually contain other entities like *Madrid* (location). A similar mistake can

be seen in *T6*. Here *Appel* was incorrectly recognized as a location instead of the surname of *Karel Appel* (person), and thus the triple represents the information to be an influence of an artist on a location, instead of between the artists.

Examples in *T2* to *T6* represent the triples and the supporting text snippets for the results of the query as depicted in Figure 4, which contains a mixture of factually correct, factually incorrect, and speculative facts. In *T2*, a relation was correctly extracted from the text, but the head entity was incorrectly recognized as 'American'. This example speaks for the need for additional work on co-reference resolution, in order to properly follow the connections in the text. A more precise triple would have been ⟨*Gorky, beInfluenceBy, Pablo Picasso*⟩.

T3 is an example in which the lack of context in the syntactic analysis of the sentence results in the assumption that the statement is true, although it is a suggestion by a specific person and therefore, not necessarily a true fact. A similar example is *T4* in which the source text is explaining a potential *influence* relation between the artists, but it cannot be directly assumed that it is a fact. These two examples illustrate that the context of the actual text might get lost during the extraction process, which may lead to erroneous facts being represented in the KG. Thus, it is important to take into account the provenance information that can help the user understand the full context for obtaining the correct information.

A different scenario is depicted in *T5*, in which the text clearly confirms the validity of the fact. One interesting observation is regarding the syntactic structure of the relation phrase - the word 'doubtless' acts as an adverb emphasizing the validity of the fact, and although it divides the relation phrase 'was influenced by', the syntactic analyzer and the canonicalization step were able to normalize the relation to a canonical form. This is also evident in the diversity of relation phrases in this sample of texts. They are expressed in different tenses, with auxiliary verbs, and sometimes spread within a more complex sentence, as seen in *T5*. Examples *T3* to *T6* illustrate the need for fact-checking in our KG. Particularly, the facts in the KG could be presented to domain experts who would be able to easily look at the information in a user-friendly manner and then proceed to investigate further to either corroborate or even contradict the triples in the automatically generated KG. We envision the easy access and scrutiny of the information stored in large text collections to be the primary use-case of this automatically generated art-historic KG.

6 Lessons Learned and Future Work

This work presented a first attempt at constructing a domain-oriented knowledge graph for the art domain in an automated fashion with Open IE techniques. Due to the noisy and heterogeneous dataset that is typical of digitized art-historic collections, we encountered challenges at various steps of the KG construction process. During the very first step, it was difficult to correctly identify the mentions of artworks (i.e. titles of paintings) in the dataset due to the noise and inherent ambiguities. This domain-specific issue needs further attention in order

to improve the quality as well as coverage of the resulting KG, as discussed in detail by previous work [23]. In addition, a co-reference resolution tool [9] could also help with the identification and linking of relevant entities.

While the Open IE approach allowed for the extraction of a wide variety of entities and relations, this led to canonicalization becoming a complicated task. We observed that existing techniques for canonicalization on generic datasets, such as CESI, do not show comparable performance for domain-specific dataset. It would be interesting to investigate if large pre-trained language models such as FastText and BERT could compete with the relatively older KG embeddings that were employed in CESI for obtaining better clusters. There are other recent works on canonicalization [24,11] that demonstrate better results and would be worth exploring further for our use case in future work. Another important aspect is the incomplete tagging of the various types of entities obtained from Open IE. Attributed yet again to the noise in the process, as well as to lack of any underlying schema, many entities could not be assigned their correct type. This task needs further exploration for the enrichment of the KG.

Moreover, we have only considered English texts in this work so far, since the existing methods show their best performance with English texts. However, our art-historic collection is comprised of multiple languages and we would like to expand the pipeline to process multi-lingual texts. Taking into account the existing limitations of the methods with domain-specific corpora, this seems to be an arduous but interesting research challenge.

With regard to the implementation of the KG pipeline, while we have so far used off-the-shelf tools and libraries like SpaCy, Stanford CoreNLP and CESI, we plan to further fine-tune them to the task of domain-specific KG construction. It will also be worthwhile to explore and evaluate the performance with other available tools such as Flair [1] and Blink [42] for entity recognition, linking and typing, as well as OpenIE [26] and MinIE [18] for the extraction of triples. The scalability of these approaches and the completeness of the resulting KG in the presence of new and expanding cultural heritage datasets is also an open research question to be looked into.

The evaluation of the art-historic KG is also a crucial task worth discussing. While we have performed a semi-automated evaluation for the first version of our KG, a more rigorous and thorough evaluation of the correctness of the facts is certainly imperative before this KG can be useful to a non-expert user (as discussed in Section 5). One way to ensure this would be to maintain the provenance and of the facts in the KG, in terms of their source document as well as their confidence measure. This could also facilitate a fair and complementary manual evaluation in terms of precision and recall which could provide further insights. For this, we plan to closely collaborate with domain experts and enlist their help in the near future.

7 Conclusion

In this work, we have presented our approach to construct an art-historic KG from digitized texts in an automated manner. We have leveraged existing Open IE tools for various stages of the KG construction process and discussed the limitations and challenges while adapting these generic tools for domain-specific datasets. We have presented these insights with the hope of encouraging interesting dialogue and further progress along these lines. While our limited initial analysis and evaluation has shown encouraging results, it has also shown clear indications towards the points of improvement for creating a more refined and comprehensive version of an art-historic KG which could be used for downstream tasks such as search and querying.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). pp. 54–59. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-4010, https://aclanthology.org/N19-4010
2. Angeli, G., Premkumar, M.J.J., Manning, C.D.: Leveraging linguistic structure for open domain information extraction. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 344–354 (2015)
3. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics **41**(5), 706–716 (2008)
4. Bonacich, P.: Power and centrality: A family of measures. American journal of sociology **92**(5), 1170–1182 (1987)
5. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E.R., Mitchell, T.M.: Toward an Architecture for Never-Ending Language Learning. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence. pp. 1306–1313 (2010)
6. 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: International Semantic Web Conference. pp. 36–52. Springer (2019)
7. Castellano, G., Sansaro, G., Vessio, G.: ArtGraph: Towards an Artistic Knowledge Graph. arXiv e-prints pp. arXiv–2105 (2021)
8. Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F., Flammini, A.: Computational fact checking from knowledge networks. PloS one **10**(6), e0128193 (2015)
9. Clark, K., Manning, C.D.: Deep reinforcement learning for mention-ranking coreference models. In: Proceedings of the 2016 Conference on Empirical Methods on Natural Language Processing (2016), https://nlp.stanford.edu/pubs/clark2016deep.pdf
10. Corro, L.D., Gemulla, R.: ClausIE: Clause-based open information extraction. Proceedings of the 22nd International Conference on World Wide Web (2013)

11. Dash, S., Rossiello, G., Mihindukulasooriya, N., Bagchi, S., Gliozzo, A.: Open knowledge graphs canonicalization using variational autoencoders. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 10379–10394. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.811, https://aclanthology.org/2021.emnlp-main.811
12. 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)
13. Ernst, P., Meng, C., Siu, A., Weikum, G.: Knowlife: A knowledge graph for health and life sciences. In: Proceedings of the 30th International Conference on Data Engineering. pp. 1254–1257. IEEE (2014)
14. Ernst, P., Siu, A., Weikum, G.: Knowlife: A versatile approach for constructing a large knowledge graph for biomedical sciences. BMC bioinformatics **16**(1), 1–13 (2015)
15. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Communications of the ACM **51**(12), 68–74 (2008)
16. Fader, A., Soderland, S., Etzioni, O.: Identifying Relations for Open Information Extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1535–1545. Association for Computational Linguistics (2011)
17. Galárraga, L., Razniewski, S., Amarilli, A., Suchanek, F.M.: Predicting Completeness in Knowledge Bases. In: Proceedings of the 10th ACM International Conference on Web Search and Data Mining. pp. 375–383 (2017)
18. Gashteovski, K., Gemulla, R., del Corro, L.: MinIE: Minimizing facts in open information extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2630–2640. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1278, https://aclanthology.org/D17-1278
19. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., et al.: Knowledge graphs. ACM Computing Surveys (CSUR) **54**(4), 1–37 (2021)
20. Huang, S., Wan, X.: AKMiner: Domain-specific knowledge graph mining from academic literatures. In: Proceedings of the International Conference on Web Information Systems Engineering. pp. 241–255. Springer (2013)
21. Hunter, J., Odat, S.: Building a Semantic Knowledge-base for Painting Conservators. In: Proceedings of the 2011 IEEE Seventh International Conference on eScience. pp. 173–180 (2011). https://doi.org/10.1109/eScience.2011.32
22. Jain, N.: Domain-Specific Knowledge Graph Construction for Semantic Analysis. In: Proceedings of the Extended Semantic Web Conference (ESWC) 2020 Satellite Events. pp. 250–260. Springer International Publishing, Cham (2020)
23. Jain, N., Krestel, R.: Who is Mona L.? Identifying Mentions of Artworks in Historical Archives. In: Proceedings of the International Conference on Theory and Practice of Digital Libraries. pp. 115–122. Springer International Publishing, Cham (2019)
24. Jiang, T., Zhao, T., Qin, B., Liu, T., Chawla, N.V., Jiang, M.: Canonicalizing Open Knowledge Bases with Multi-Layered Meta-Graph Neural Network. CoRR **abs/2006.09610** (2020), https://arxiv.org/abs/2006.09610
25. Kejriwal, M.: Domain-specific knowledge graph construction. Springer (2019)

26. Kolluru, K., Adlakha, V., Aggarwal, S., Mausam, Chakrabarti, S.: OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 3748–3761. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.306, https://aclanthology.org/2020.emnlp-main.306
27. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., 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)
28. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014), http://www.aclweb.org/anthology/P/P14/P14-5010
29. Nickel, M., Rosasco, L., Poggio, T.: Holographic embeddings of knowledge graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 30 (2016)
30. Oldman, D., Labs, C.: The CIDOC Conceptual Reference Model (CIDOC-CRM): PRIMER. CIDOC-CRM official web site (2014)
31. Oramas, S., Espinosa-Anke, L., Sordo, M., Saggion, H., Serra, X.: Information extraction for knowledge base construction in the music domain. Data and Knowledge Engineering **106**, 70–83 (2016). https://doi.org/10.1016/j.datak.2016.06.001, https://www.sciencedirect.com/science/article/pii/S0169023X16300416
32. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab (November 1999), http://ilpubs.stanford.edu:8090/422/, previous number = SIDL-WP-1999-0120
33. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020), https://nlp.stanford.edu/pubs/qi2020stanza.pdf
34. Rudnik, C., Ehrhart, T., Ferret, O., Teyssou, D., Troncy, R., Tannier, X.: Searching news articles using an event knowledge graph leveraged by Wikidata. In: Companion Proceedings of The 2019 World Wide Web Conference. pp. 1232–1239 (2019)
35. Shin, J., Wu, S., Wang, F., De Sa, C., Zhang, C., Ré, C.: Incremental Knowledge Base Construction using Deepdive. In: Proceedings of the VLDB Endowment International Conference on Very Large Data Bases. vol. 8, p. 1310 (2015)
36. Tietz, T., Waitelonis, J., Zhou, K., Felgentreff, P., Meyer, N., Weber, A., Sack, H.: Linked Stage Graph. In: SEMANTICS Posters&Demos (2019)
37. Van Hooland, S., Verborgh, R.: Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata. Facet publishing (2014)
38. Vashishth, S., Jain, P., Talukdar, P.: CESI: Canonicalizing open knowledge bases using embeddings and side information. In: Proceedings of the 2018 World Wide Web Conference. pp. 1317–1327 (2018)
39. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledge base. Communications of the ACM **57**(10), 78–85 (2014)
40. Vsesviatska, O., Tietz, T., Hoppe, F., Sprau, M., Meyer, N., Dessì, 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)

41. Wu, H., Liu, S.Y., Zheng, W., Yang, Y., Gao, H.: PaintKG: The painting knowledge graph using biLSTM-CRF. In: Proceedings of the 2020 International Conference on Information Science and Education (ICISE-IE). pp. 412–417 (2020). https://doi.org/10.1109/ICISE51755.2020.00094
42. Wu, L., Petroni, F., Josifoski, M., Riedel, S., Zettlemoyer, L.: Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6397–6407. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.519, https://aclanthology.org/2020.emnlp-main.519
43. Yates, A., Banko, M., Broadhead, M., Cafarella, M.J., Etzioni, O., Soderland, S.: Textrunner: Open Information Extraction on the Web. In: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT). pp. 25–26 (2007)
44. Yuan, J., Jin, Z., Guo, H., Jin, H., Zhang, X., Smith, T., Luo, J.: Constructing biomedical domain-specific knowledge graph with minimum supervision. Knowledge and Information Systems **62**(1), 317–336 (2020)
45. Zhao, X., Xing, Z., Kabir, M.A., Sawada, N., Li, J., Lin, S.W.: HDSKG: Harvesting domain specific knowledge graph from content of webpages. In: Proceedings of the 24th International Conference on Software Analysis, Evolution and Re-engineering (SANER). pp. 56–67. IEEE (2017)