

Knowledge Extraction for Art History: the Case of Vasari's *The Lives of The Artists* (1568)

Cristian Santini^{1,2}, Mary Ann Tan^{1,2}, Oleksandra Bruns^{1,2}, Tabea Tietz^{1,2},
Etienne Posthumus¹ and Harald Sack^{1,2}

¹FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Karlsruhe, Germany

²Karlsruhe Institute of Technology, Institute AIFB, Karlsruhe, Germany

Abstract

Knowledge Extraction (KE) techniques are used to convert unstructured information present in texts to Knowledge Graphs (KGs) which can be queried and explored. Despite their potential for cultural heritage domains, such as Art History, these techniques often encounter limitations if applied to domain-specific data. In this paper we present the main challenges that KE has to face on art-historical texts, by using as case study Giorgio Vasari's *The Lives of The Artists*. This paper discusses the following NLP tasks for art-historical texts, namely entity recognition and linking, coreference resolution, time extraction, motif extraction and artwork extraction. Several strategies to annotate art-historical data for these tasks and evaluate NLP models are also proposed.

Keywords

Knowledge Extraction, Art History, Cultural Heritage, NLP

1. Introduction

To understand the meaning of works of art, art historians investigate and analyze not only the physical properties and subject matter, but the historical context of their creation. A rich source of historical knowledge is evident in textual documents that contain artifact descriptions, historical memoirs, letters, and biographies, such as Giorgio Vasari's *The Lives of The Artists* (1568). Nowadays, the information contained in these historical texts is not only of interest for art researchers, but also for computer scientists who want to fully unlock the knowledge contained therein with the help of Natural Language Processing (NLP).


NLP techniques for Knowledge Extraction (KE), like Entity Recognition (ER), Entity Linking (EL) and Relation Extraction (RE), are used to convert unstructured information present in texts to Knowledge Graphs (KGs) which can be queried and explored. However, domain specific resources, such as art-historical texts, pose a number of challenges when these tasks are conducted. For example, most NLP techniques don't take into account the problems specific to historical texts: the complexity of historic languages, the variety of the types of documents (press articles, letters, diaries, etc.), the specific pragmatics of the authors, and the possible noise

Qurator 2022: 3rd Conference on Digital Curation Technologies, September 19-23, 2022, Berlin, Germany

✉ cristian.santini@fiz-karlsruhe.de (C. Santini); ann.tan@fiz-karlsruhe.de (M. A. Tan);
oleksandra.bruns@fiz-karlsruhe.de (O. Bruns); tabea.tietz@fiz-karlsruhe.de (T. Tietz);
etienne.posthumus@partners.fiz-karlsruhe.de (E. Posthumus); harald.sack@fiz-karlsruhe.de (H. Sack)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

present in the input due to data processing techniques such as Optical Character Recognition (OCR). Moreover, there is a lack of domain-specific resources, especially for art-historical data, to train Machine Learning (ML) models to extract or classify specific information for art-historians. For this reason, an identification of challenges specific to KE on art-historical texts is needed.

This paper discusses a series of challenges which emerged when dealing with KE on *The Lives of The Artists*, a collection of artists' biographies written by the Italian painter Giorgio Vasari in the XVIth century. This book is still one of most authoritative bibliographic references for Art History from the Gothic to the Mannerist era and to extract a knowledge graph from this work would entail the possibility for researchers to explore and query information related to European artists from the XIIIth to the XVIth century, their creative endeavours and the dynamics of their conservation, which is still not available in general-purpose KGs, such as Wikidata. Moreover, this digitization process would allow, from an historiographical perspective, to compare statements made by Vasari with those availed by contemporary art historians. For this reason, a well-designed evaluation dataset which enables a fair comparison of different NLP tools on this text is needed. Scope of this work is the analysis of the issues involved in the curation of art-historical datasets and the presentation of potential solutions for their annotation.

The challenges of KE for Art History presented in this paper are related to the following NLP tasks: A) the recognition and resolution of entities associated with various concepts, i.e. ER and EL; B) the resolution of coreferences between entities and pronouns in long texts; C) the extraction and normalization of temporal information; D) the extraction of artwork-related information; E) the possibility of considering, during the evaluation, multiple correct annotations in the ground truth. More specifically, D) does not only concern the identification of contextual information, such as used materials, locations or dates, but also to the extraction or classification of artworks' content. In this context, the identification of so-called motifs, i.e. iconographic themes which involve one or multiple subjects, is of course of interest.

The contribution of this paper is the discussion of these aforementioned challenges along with the delineation of the possible solutions that NLP practitioners can adopt to annotate art-historical data and create KE resources for this domain. In order to provide examples to motivate our strategies, we used as case-study excerpt extracted from *The Lives of The Artists* by Vasari (1568). The remainder of the paper is structured as follows. Section 2 presents related work on KE and the Digital Humanities. Section 3 describes the main challenges for KE on art-historical texts by using as case study *The Lives of The Artists*. Section 4 concludes the paper.

2. Related Work

A general overview of the variety of knowledge extraction tasks which have been carried out for cultural heritage domains is provided in [1]. This work describes the scenarios in which NLP techniques are applied (data processing, knowledge extraction, metadata extraction) and a series of general challenges for cultural heritage data. However, an overview of the specific challenges for specific domains is lacking. A recent survey on the challenges of ER and EL for historical documents is provided in [2], where the authors list four main issues: the variety of documents (newspapers, letters, memoirs, books), the presence of noise in the input data due to

data processing techniques (such as OCR), the problem of old historical language varieties and the lack of standardized resources. In fact, as motivated by this study, the lack of standardized benchmarks for historical data pushes researchers to focus their effort either on specific tasks, such as entity recognition and linking, or on specific document types, such as old historical press data or literary texts.

One of the main KE tasks carried out in the Digital Humanities is EL, due to the importance of correctly resolving mentions to entities to a KG. [3] utilized DBpedia Spotlight to disambiguate named entities on the Bentham Corpus. In [4], three third-party entity extraction tools were evaluated on a case study based on the descriptive fields of the Smithsonian Cooper-Hewitt National Design Museum in New York. With respect to these studies, common limitation is the fact that third-party tools are used on English-only texts and the creation of publicly available datasets to train and evaluate domain-specific models is not considered. [5] evaluated an EL architecture specifically designed for recognizing and linking mentions to persons in French literary criticism texts and scientific essays from the 19th and early 20th centuries. In [6] the problem of entity linking on multilingual historical press articles with OCR errors was addressed by evaluating an End-to-End EL model on a domain-specific corpus: the NewsEye Benchmark Dataset¹. However, a main limitation of these studies is the fact that the proposed approaches consider only a restricted set of entities, either persons [5] or proper nouns [6], so these systems may not work well in the art-history domain, where significant variations in vocabulary and lexical and syntactic structures are present.

Named Entity Recognition (NER), Coreference Resolution and Event Detection was addressed in [7] by creating a benchmark from literary works extracted from project Gutenberg, containing 100 different English-language novels. However, this resource reuses the guidelines of a general-domain dataset, i.e. ACE². Specific to this dataset is the inclusion of common nouns (*boy*, *kitchen*) and nested structures (such as *[[the cook]'s sister]*). The annotation and normalization of temporal information, i.e. Temporal Tagging, for historical documents was proposed in [8], where a multilingual corpus of historical documents extracted from Wikipedia with annotated temporal information is presented. As main challenges for temporal tagging on these documents, the authors mention the importance of context, underspecificity, the portability of domain-specific guidelines to contemporary texts and the consideration of language specific differences. However, since the corpus consists of Wikipedia articles, this study does not take into account the problem of historical lexical variations used to express time.

Few studies have investigated EL and, more generally, knowledge extraction techniques for the Art History domain. [9] proposed an EL corpus for creative works extracted from contemporary texts. In their study, the authors focused on the annotation of nested entities and their impact on the evaluation of three general-purpose EL tools: Dbpedia Spotlight, AIDA and Recognize. Despite its focus on the challenges of textual data containing mentions to creative works, this work does not take into consideration further challenges of the annotation, such as the definition of domain-specific tagsets or the inclusion of multiple correct annotations for the same entity.

¹<https://zenodo.org/record/4573313#.YtUp8HZByn4>

²<https://catalog.ldc.upenn.edu/LDC2006T06>

3. Challenges of Knowledge Extraction in *The Lives of The Artists*

In this section, challenges of the following KE tasks on art-historical texts will be discussed: entity recognition, entity linking, coreference resolution, time extraction, motif extraction and artwork extraction. These challenges emerged as part of a preliminary research on the creation of a KE pipeline for Vasari's *The Lives of The Artists*. The goal of the aforementioned research is to apply a series of NLP techniques to art-historical texts in order to verify their potential in complementing the information present in existing KGs. However, a number of challenges specific to art-historical texts were found. In each of the following subsections, potential strategies to overcome these challenges and create task-specific annotations are presented.

3.1. Entity Recognition

One of the first steps in a Knowledge Extraction pipeline is the identification of entity mentions inside a text, namely Entity Recognition (ER). A problem to be faced when dealing with art-historical texts is related to the fact that entities of interest are also expressed as common nouns: materials from which artifacts are made, techniques, styles, subjects, etc. For this challenge, fine-grained tagsets, such as the one from the recently released MultiNERD [10] can be reused. This tagset contains entity types applicable to both named entities and common nouns which occur in art-historical data, such as Persons (PER), Organizations (ORG), Locations (LOC), Event (EVE), Animal (ANIM), Plant (PLANT) and Mythological entity (MYTH). To extend the annotation to cultural products and related information, this tagset can be extended with the following entity types:

- Artifact (ART): both tangible, such as books (*Old Testament*), artworks, instruments; and intangible, such as subject matters (*philosophy*), languages, iconographies (*Mary with the Child*).
- Material (MAT): materials from which artifacts can be made (*wax*)
- Technique (TECH): technique with which an artifact is made (*fresco*)

But meanwhile the [Florentines]_{PER}, hearing in the year 1515 that [[Pope]_{PER} Leo X]_{PER} wished to grace his native [city]_{LOC} with his presence, ordained for his [reception]_{EVE} extraordinary [festivities]_{EVE} and a sumptuous and magnificent [spectacle]_{EVE}, with so many [arches]_{ART}, [façades]_{ART}, [temples]_{ART}, [colossal figures]_{ART}, and other [statues]_{ART} and [ornaments]_{ART}, that there had never been seen up to that time anything richer, more gorgeous, or more beautiful; for there was then flourishing in that [city]_{LOC} a greater abundance of fine and exalted [intellects]_{PER} than had ever been known at any other period.

E 1: Annotation considering common nouns, proper nouns, noun phrases and nested entities.

3.2. Entity Linking

A further step in KE is the resolution of identified entity mentions by using a KG, i.e. EL. One of the first challenges that should be solved is to find a KG with extensive coverage of the entity types mentioned in Section 3.1. Wikidata is a strong candidate due to its wide coverage.

However, by using a general-purpose KG, entities from historical texts, such as those mentioned in Vasari’s text, may not be found, and therefore they are to be labelled as NIL. Moreover, another challenge is related to historical languages and the presence of lexical variations to define known entities. For example, what Vasari often mentions as *Greek manner* refers to the Byzantine art. In addition, figures of speech such as metonymy are to be considered. E 2 shows how the surface form *Fontainebleau* is resolved by a link to the implicitly mentioned palace and not to the geographical place. In E 2 each entity is resolved by using the corresponding title of the English Wikipedia page.

The [work]*Creative_Work* is now in the [collection]*Collection_(museum)* of [[King]*King Francis* of [France]*France*]*Francis_I_of_France*, at [Fontainebleau]*Palace_of_Fontainebleau*.

E 2: Annotation showing how entities are resolved.

3.3. Coreference Resolution

Art-historical texts may often be long documents, such as biographies or correspondence. Therefore, coreferences between previously explicitly mentioned entities and personal pronouns should be taken into account. E 3 shows a sentence in Vasari’s text where the first clause does not have as subject an explicit mention to an entity but only a personal pronoun. For coreference resolution, a possible strategy is to annotate the surface forms of the entities from the entity recognition dataset as well as personal pronouns, as previously performed in [7] for non-contemporary novels. Following the guidelines proposed in the same work, nested honorifics (*Mr.*) and titles (*Pope*) are excluded from the coreference resolution task.

[He]₁ also restored the [right arm]₂ of the ancient [Laocoon]₃, which had been broken off and never found, and [Baccio]₁ made one of the full size in [wax]₄, which so resembled the ancient [work]₃ that [it]₂ showed how [Baccio]₁ understood [his]₁ art; and this [model]₅ served [him]₁ as a [pattern]₆ for making the whole [arm]₇ of [his]₁ own [Laocoon]₈.

E 3: Annotation considering coreference.

3.4. Time Resolution

The extraction of temporal information is also relevant to capture further information about the contexts of artworks and artists’ lives. One of the challenges of time extraction in historical documents is presented by non-specific time expressions, such as references to centuries (*6th century*), historical periods (*The Baroque era*) and fuzzy time expressions (*By the year 1420*). For general time expressions, annotation frameworks such as TimeML [11] have been proposed to normalize them; moreover, the use of information inside KG, such as Wikidata, allow to extract dates of historical periods using `start_date` and `end_date` properties. E 4 shows how information related to historical periods can be annotated by using TIMEX3 INTERVAL [12].

```
<TIMEX3INTERVAL earliestBegin="1464-09-16" latestBegin="1464-09-16"
earliestEnd="1471-07-26" latestEnd="1471-07-26"><TIMEX3 tid="t1"
type="DURATION" value="P2504D">In the time of Pope Paul II</TIMEX3>
</TIMEX3INTERVAL>
```

E 4: Example of annotated and normalized time expression.

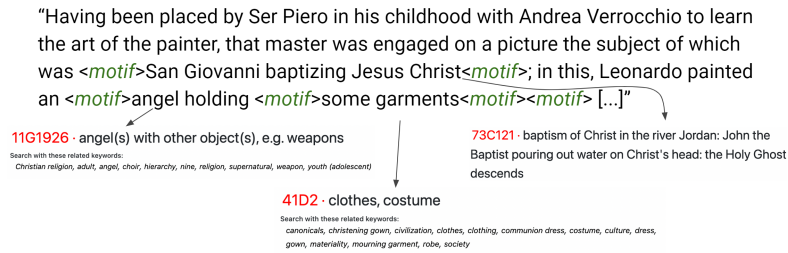


Figure 1: Example of annotation with nested motifs.

However, a further complexity found in Vasari is posed by references to calendar dates which use mentions of popular events, such as festivities. For example, in this book the date of birth of Raffaello is mentioned as the *Good Friday* of the year 1483. To normalize these expressions, we can rely on the most consolidated opinion among art-historians, which are using the Gregorian Calendar to indicate the Good Friday of 1483 as March 28.

3.5. Motif Extraction

Analysis of artistic themes or motifs in artworks is a vital endeavor for art historians. Motifs can be either named entities (*S. Sebastian*), long noun phrases (*The Feast in the House of Simon the Pharisee*) or common nouns (*vase*). It is clear that these types of entities cannot be resolved with an encyclopedic KG such as Wikidata, since it is not designed for iconographic subjects. Moreover, for the extraction of this information two tasks should be carried out: a) the recognition of sequences of words which describe iconographic motifs, and b) the identification of these motifs by using an appropriate classification system.

This can be addressed by modelling the problem as a specific *sequence labelling* task where motifs are identified inside a text and annotated through the ICONCLASS³ classification system [13]. This task is distinguished from ER and EL since the relations between multiple entities and specific events should be considered. Moreover, another challenge is related to the fact that the same motif can be described with nested ICONCLASS notation, as shown in Figure 1.

3.6. Artwork Extraction

The identification of artworks is crucial when dealing with art-historic data. However, artistic works are not always mentioned explicitly in historical texts. In fact, complex lexical variations and long descriptions are often used by non-contemporary art-historians to refer to artists' endeavours. Figure 2 shows an example of two paintings referenced by Vasari: despite the fact that the explicit title is missing, the artwork references can still be resolved by a human annotator to a unique entry in a Knowledge Base (KB) (in this case Wikipedia). For this reason, further KE tasks should be considered, besides those previously mentioned: a) recognition of explicit as well as implicit artwork references; and b) resolution of these descriptions by using a unique identifier present in an external resource, such as a KB or a KG.

³www.iconclass.org

In the tramezzo of the Ognissanti, by the door that leads into the choir, he painted for the Vespucci a **S. Augustine** in fresco.

Piero painted, for the elder Filippo Strozzi, a **picture with little figures of Perseus delivering Andromeda from the Monster**, in which are some very beautiful things



Figure 2: Example of recognition and linking of artworks' references: the second sentence (from the top) shows a long description linked to a specific painting.

Sequences of words referring to artworks are often triggered by verb phrases which are referred to the activity of an artist (*painting*). Moreover, for artwork resolution, the previously described steps of ER and EL, coreference resolution, time extraction and motif extraction should come into play. The second example (from the top) in Figure 2 shows how to correctly interpret the creator of a work a personal pronoun has to be resolved. Finally, artwork descriptions allow us to extract further contextual information related to an artifact, such as its location and commissioner, or information related to its depicted motifs, which might complement the already existing knowledge about an artwork.

3.7. Evaluation

With the KE tasks identified in the preceding sections, a comprehensive evaluation strategy is necessary. In this context, the pragmatics behind the evaluation of NLP tools should take into consideration the potential ambiguity of several linguistic expressions related to cultural concepts. As an example, nested entities are relevant when dealing with ER and iconographies, since these often consists of sets of elements, as in *[[Mary] with the [Child]]*.

Another problem is related to the possibility of multiple correct answers for entity resolution systems. As an example, artwork descriptions in Vasari's text which mention the deity of Love may be resolved by two different annotators by using the Wikidata entities Eros (Q121973) or Cupid (Q5011), being both alias of the same entity. With respect to this issue, leveraging annotators' disagreement can be beneficial, as discussed in [14].

A further challenge applies to motifs: for example, the subject of the *Last Supper* can be classified in ICONCLASS using 2 entries: 73D2, which identifies the biblical episode, and 73D24,

which is a subclass of the first and identifies the specific scene. While a human annotator can try to use always the most precise class in the ICONCLASS hierarchy, an automatic system may classify a motif with a superclass or subclass of the ICONCLASS code in the ground-truth. For this, evaluation measures for hierarchical classification, such as those proposed in [15], can be considered.

4. Conclusion and Future Work

Art-historical texts are part of our cultural heritage and as such can offer additional insights on how ideas developed around artistic subjects and which conditions determined the evolution of artistic tendencies. In this paper a series of challenges of KE for Art History are presented. The takeaway lesson is that, to model a specific domain, the concepts introduced and the limitations of existing general approaches should be carefully considered. Moreover, complex linguistic features, ambiguity and historical ideas that are not anymore part of our cultural frame pose challenges to NLP models. In future work, a fully annotated sample dataset extracted from *Lives of The Artists* will be provided to enable a valid comparison of KE tools for art-historical texts, as well as complete and detailed guidelines for the annotation and the evaluation of the proposed KE tasks.

References

- [1] C. Sporleder, Natural Language Processing for Cultural Heritage Domains, *Language and Linguistics Compass* 4 (2010) 750–768. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2010.00230.x>. doi:10.1111/j.1749-818X.2010.00230.x, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-818X.2010.00230.x>.
- [2] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet, Named Entity Recognition and Classification on Historical Documents: A Survey, 2021. URL: <http://arxiv.org/abs/2109.11406>, arXiv:2109.11406 [cs].
- [3] P. Ruiz, T. Poibeau, Mapping the Bentham Corpus: Concept-based Navigation, *Journal of Data Mining and Digital Humanities Atelier Digit_Hum* (2019). URL: <https://hal.archives-ouvertes.fr/hal-01915730>. doi:10.46298/jdmdh.5044, publisher: Episciences.org.
- [4] S. van Hooland, M. De Wilde, R. Verborgh, T. Steiner, R. Van de Walle, Exploring entity recognition and disambiguation for cultural heritage collections, *Digital Scholarship in the Humanities* 30 (2015) 262–279. URL: <https://doi.org/10.1093/llc/fqt067>. doi:10.1093/llc/fqt067.
- [5] C. Brando, F. Frontini, J.-G. Ganascia, REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets, *Complex Systems Informatics and Modeling Quarterly* (2016) 60–80. URL: <https://csimq-journals.rtu.lv/article/view/csimq.2016-7.04>. doi:10.7250/csimq.2016-7.04, number: 7.
- [6] E. Linhares Pontes, L. A. Cabrera-Diego, J. G. Moreno, E. Boros, A. Hamdi, A. Doucet, N. Sidere, M. Coustaty, MELHISSA: a multilingual entity linking architecture for historical

- press articles, *International Journal on Digital Libraries* 23 (2022) 133–160. URL: <https://doi.org/10.1007/s00799-021-00319-6>. doi:10.1007/s00799-021-00319-6.
- [7] D. Bamman, S. Popat, S. Shen, An annotated dataset of literary entities, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2138–2144. URL: <https://aclanthology.org/N19-1220>. doi:10.18653/v1/N19-1220.
 - [8] J. Strötgen, T. Bögel, J. Zell, A. Armiti, T. V. Canh, M. Gertz, Extending HeidelTime for Temporal Expressions Referring to Historic Dates, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 2390–2397. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/849_Paper.pdf.
 - [9] A. M. Brasoveanu, A. Weichselbraun, L. Nixon, In Media Res: A Corpus for Evaluating Named Entity Linking with Creative Works, in: *Proceedings of the 24th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Online, 2020, pp. 355–364. URL: <https://aclanthology.org/2020.conll-1.28>. doi:10.18653/v1/2020.conll-1.28.
 - [10] S. Tedeschi, R. Navigli, MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation), in: *Findings of the Association for Computational Linguistics: NAACL 2022*, Association for Computational Linguistics, Seattle, United States, 2022, pp. 801–812. URL: <https://aclanthology.org/2022.findings-naacl.60>.
 - [11] R. Saurí, J. Moszkowicz, B. Knippen, R. Gaizauskas, A. Setzer, J. Pustejovsky, TimeML Annotation Guidelines Version 1.2.1 (2006).
 - [12] E. Kuzey, J. Strötgen, V. Setty, G. Weikum, Temponym Tagging: Temporal Scopes for Textual Phrases, in: *Proceedings of the 25th International Conference Companion on World Wide Web, WWW ’16 Companion*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2016, pp. 841–842. URL: <https://doi.org/10.1145/2872518.2889289>. doi:10.1145/2872518.2889289.
 - [13] L. D. Couprie, Iconclass: an iconographic classification system, *Art Libraries Journal* 8 (1983) 32–49. URL: <https://www.cambridge.org/core/journals/art-libraries-journal/article/abs/iconclass-an-iconographic-classification-system/DD119669E055893AB632E5C7CE6FF417>. doi:10.1017/S0307472200003436, publisher: Cambridge University Press.
 - [14] L. Aroyo, C. Welty, Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation, *AI Magazine* 36 (2015) 15–24. URL: <https://ojs.aaai.org/index.php/aimagazine/article/view/2564>. doi:10.1609/aimag.v36i1.2564, number: 1.
 - [15] A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, I. Androutsopoulos, Evaluation measures for hierarchical classification: a unified view and novel approaches, *Data Mining and Knowledge Discovery* 29 (2015) 820–865. URL: <https://doi.org/10.1007/s10618-014-0382-x>. doi:10.1007/s10618-014-0382-x.