Viewsari: Enabling New Perspectives on the Renaissance with a Knowledge Graph of Giorgio Vasari's The Lives

Sarah Rebecca Ondraszek[1,2,*]

[1]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Germany
[2]Institute of Applied Informatics and Formal Description Methods (AIFB) of KIT, Karlsruhe, Germany

Abstract

In the digital humanities, semantic technologies have been recognized as providing the necessary bits and pieces to represent the complex and often ambiguous nature of humanities data. Despite this growing interest, a lack of practical frameworks for modeling the complex, usually multifaceted and multilingual, historical sources remains. In this paper, we present Viewsari, an ongoing Ph.D. project aiming to build a knowledge graph based on Giorgio Vasari's *Lives of the Most Excellent Painters, Sculptors, and Architects* (1568), referred to as *The Lives*. This collection of biographies of important Renaissance artists, recounting tales of their lives and describing their artistic styles and works, is widely regarded as the first modern work of art history. With it, Vasari shaped the canon of the Italian Renaissance.

The Viewsari project draws on knowledge extraction and aims to contextualize content from different editions of Vasari's *The Lives*, addressing the challenges of working with complex, multilingual historical texts. Situated at the intersection of digital humanities and the Semantic Web, it demonstrates how modular, pattern-driven ontology development, leveraging Ontology Design Patterns and the eXtreme Design methodology, can support the structured representation and exploration of information across different editions and linguistic versions. The central goal is to generalize the Viewsari framework to match similar challenges, i.e., enriching and interconnecting textual sources in different domains.

Keywords

Digital humanities, knowledge graphs, ontologies, knowledge extraction

1. Introduction

As an interdisciplinary field of study, the digital humanities (DH) encompasses domains such as art history, performative arts, and literary studies, all of which provide heterogeneous data. Thus, it represents a challenging domain for the application and evaluation of semantic technologies [1, 2]. Their application in the DH, however, often results in isolated or overly complex solutions that lack interoperability and reusability. This partly stems from the heterogeneous nature of the data, which frequently compels researchers to create ad hoc solutions instead of modularized approaches [3, 4].

This Ph.D. project lies at the intersection of DH and the Semantic Web. With it, we explore how semantic technologies can enable structured, interoperable, and reusable knowledge representations to support the modeling and analysis of complex and frequently multilingual (historical) texts. To this end, we incorporate ontology engineering, knowledge graph (KG) construction, and knowledge extraction using natural language processing (NLP) tools and state-of-the-art (SOTA) large language models (LLMs).

The chosen case study to demonstrate these aspects is Giorgio Vasari's Renaissance literary work *The Lives of the Most Eminent Painters, Sculptors, and Architects* (The Lives) [5].[1] As a collection of biographies of prominent artists of the time, it is considered one of the founding works of art history as a discipline. Over the past 450 years, various efforts have contributed to its analysis, including scholarly

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

*Corresponding author.

sarah-rebecca.ondraszek@fiz-karlsruhe.de (S. R. Ondraszek)
0009-0003-7945-6704 (S. R. Ondraszek)

[1] Original title: *Le vite de' più eccellenti pittori, scultori e architettori.* Vasari originally published the first version in 1550 and later expanded and revised it, resulting in the 1568 edition, which includes additional biographies.

editions, commentaries, annotations, and reinterpretations. These efforts have typically followed the tradition of "close reading," focusing on detailed, line-by-line interpretations [6, 7].

Recent years have shown a growing interest in Franco Moretti's concept of scaling literary analysis across larger corpora of texts to a 'bird-perspective', so-called 'distant reading', analyzing patterns using computational methods to uncover broader connections [8].

Despite the shift, to this day, *The Lives*, among other seminal works, remains a source of unstructured text. However, a structured representation is necessary to represent the multifaceted content of many sources, be it historical, scientific, or of fantastic matter. In Vasari's case, this can open the biographies to interdisciplinary research questions, facilitate navigation and interpretation of the web of relationships and contexts described, and possibly allow for new insights and perspectives on the data from both a close and distant reading perspective. Entity extraction and linking lay the foundation for this construction, enabling machine-understandable, interoperable representations [9].

We introduce this project as two-fold: One part is the overall Viewsari project (referred to simply as *Viewsari*), which symbolizes the framework and entire process – *View*sari, because it intends to support an experience of the Renaissance through Vasari's eyes. The other part is the Viewsari KG as the beating heart of this Ph.D. project. Central aspects of the Viewsari KG are the formal representations of contextual implications for and in (historical) social networks and the representation of subjective claims about situations to explicitly address digital hermeneutics and context sensitivity as critical aspects in the interpretation of information [10, 11, 12]. Furthermore, with the knowledge extraction pipeline behind Viewsari, we investigate the use of NLP and LLMs to extract named entities from texts, especially those with ambiguous references or mere textual descriptions, supported by the iconographic interpretation model by Panofsky in the case of artworks [13, 14].[2]

Overall, our goal is to create a generalizable framework that can be used outside this Ph.D. project and in other domains (outside of DH) to transform unstructured texts into ontology-based KGs. Results shall be provided in a comprehensible interface, with a focus on the formal representation of content and text-related information across various versions, whether linguistic, structural, or editorial.

Problem Statement. We identify the key challenge to be a lack of repeatable methods for extracting structured information from these sources and linking it to existing databases. Especially due to linguistic variations and ambiguities, this remains a significant hurdle. This also concerns recognizing entities from their description and the treatment of long-tail cultural heritage data: lesser-known artists or vague artwork references through textual descriptions, which are often crucial for humanities research but not represented in standard knowledge bases, so-called out-of-knowledge-base (OOKB) entities [15]. Traditional NLP techniques often fail in these domains. However, recent innovations in LLMs promise to bridge this gap [9, 16, 17].

Furthermore, the creation of consistent, reusable, and comprehensible formal representations that find application beyond a single research project (especially in DH) remains a challenge [1, 4].

The challenges frequently encountered in extracting knowledge from historical texts and modeling their content, such as ambiguous terminology, interpretive knowledge, multilinguality, lengthy descriptions instead of direct references, and the need for detailed provenance, are also relevant in other domains. The same holds true for knowledge representations. This points to the need for generalizable solutions.

Importance. *The Lives* provides a rich and complex data set. It has had a lasting influence on art historical research and sets well-defined boundaries for a contained scope, making it a valuable proof-of-concept for the domain. Although the scope allows for the focused development and evaluation of proposed methodologies, the dataset also presents significant challenges: the extraction of unknown entities and the representation of historical context. This includes fine-tuning prompts for LLMs and experimenting with potential models for entity matching, especially with unseen entities [18]. To support digital hermeneutics and source criticism, Viewsari offers a structured representation. Moreover, a key feature is the formal representation of provenance information, contextual implications, and

[2] The sources for the Viewsari project can be found via https://github.com/ISE-FIZKarlsruhe/viewsari.

historical relatedness [10, 11, 19]. Additionally, the underlying ontology development process serves as a proof-of-concept for overcoming inconsistent modeling approaches and the difficulty of aligning formal representations with interpretively layered data, aiming to develop transferable best practices that researchers can apply in other DH-related contexts.

Solving this problem has value for multiple communities beyond the humanities. As a reproducible knowledge extraction and KG construction workflow, the methodology developed in Viewsari addresses the aforementioned challenges, which can also be prevalent in applications such as scientific knowledge representation, the representation of complex processes and their provenance in experiments, or social/biographical networks [20, 21]. This positions the Viewsari methodology with broader relevance and motivates its potential generalization.

2. Related Work

Shaping the way scholars conduct research in the humanities, digital methods brought new perspectives, unlocking new perceptions of what was previously hidden in source material at a larger scale [8, 12]. The following section provides an overview of related work in the context of the Viewsari project, focusing on knowledge extraction with LLMs, related challenges, and ontology engineering, particularly in the DH. It also explores cross-domain approaches to the diverse challenges mentioned in the previous section.

2.1. Knowledge Extraction with Large Language Models

Information extraction techniques such as named entity recognition (NER) and entity linking provide a way to access knowledge stored in unstructured, text-based material, and for transforming these sources into structured representations [12].

Despite the inherent heterogeneity of cultural heritage data, a commonly agreed-upon set of information is relevant to many research questions [22, 23]. Named entities, which include people, places, and historical events, are a crucial part of this. In early attempts to extract these, as shown in a survey by Sporleder, the pipelines for NER often involved classification tasks that required domain-specific training with annotated data that was frequently unavailable, and therefore, the models performed poorly.

Recent work with SOTA transformer-based technologies, such as BERT- or GPT-based models, showed that improved extraction would be possible in a zero-shot setting [12, 25]. The same applies to text-to-KG approaches, which mitigate semantic parsing to transform unstructured sources to structured graph representations, using abstract meaning representations [26]. The application of these techniques in specific domains, such as art history, is still being explored, and thus continues to pose problems in the area of artwork or motif recognition and linking from unstructured, descriptive passages [9, 27].

2.1.1. Ambiguities and Descriptive Passages in Entity Extraction

Extracting vaguely referenced entities, even with SOTA approaches and LLMs, remains a challenge. In particular, this concerns references via descriptive passages or non-named references. Furthermore, non-standardized terminology and variation in writing styles can complicate the extraction of relevant entities from unstructured texts [20]. Recent work like AI4DiTraRe [28] has shown that research data frequently lacks terminological consistency and contains nested concepts. An example is the ambiguous references to parameters in studies, such as ‘Age’, which might appear as ‘AGE’, ‘Age of sample’, or ‘age (years)’, leading to difficulties in extraction as a singular entity. As a normalization framework, AI4DiTraRe proposes an LLM-based pipeline. Similarly, Viewsari applies prompt-based pipelines with LLMs to extract and disambiguate named entities from Vasari’s text, given implicit references and cases when no canonical label is present.

2.1.2. Out-of-Knowledge-Base Entities

Advances in KG construction have addressed the challenge of treating OOKB with graph neural networks (GNNs), generating embeddings for unseen entities [15]. Projects like CHAD-KG exemplify the integration with cultural heritage data, providing a reproducible pipeline to accommodate OOKB entities with mapping rules to normalize and create IRIs for input data [29]. To expand on these works, within this Ph.D. project, in addition to the extraction of OOKB entities, a focus is on formally describing their provenance and uncertainty in an ontology.

2.2. Ontology Engineering in the Digital Humanities

Over the years, several methodologies have emerged to facilitate ontology development. One of these is XD, a dynamic approach to ontology engineering. Different, iterative phases guide the process in close consultation with domain experts. All steps focus on reusability and interoperability of developed formalizations. At the heart of this are ODPs, modular solutions for formalizations, addressing recurring design problems in ontologies, which can be shared across domains and reused like templates [30, 31, 32]. The advantages of pattern-based approaches have been explored in several works, highlighting aspects such as a lower overall error rate or higher understandability [31].

In the domain of cultural heritage, XD and patterns have been applied, e.g., in the ArCo KG [33], which integrates Italian resources. However, the broader adoption of the practices remains uncharted territory, and there is no overarching standard for connecting diverse ontologies in the DH [10, 34]. In art history, patterns have been utilized to encode the complex interpretation process of artworks [13] or digital collections of artworks [35]. Additionally, KGs have been highlighted for encoding contextual implications for and in (historical) social networks [19, 36, 37]. This approach relates to a similar endeavor from Shimizu et al. (2023), in which they introduce the ongoing work of building a library for modular ontology design (MODL), aiming to abstract and harmonize patterns across different ontologies. Similarly, in Viewsari, the goal is to reduce fragmentation and bridge the gap between pattern-based and text-based contextual annotations in a structured KG.

2.3. Cross-Domain Approaches and Shared Challenges

Inherent in different types of data, the challenges faced when extracting knowledge from historical texts and modeling the outcome and its underlying process are applicable to other domains.

Existing research in the biomedical domain tackles similar issues, including provenance and process modeling. As can be seen in ontologies like Provenance, Authoring and Versioning (PAV) [38] and extensions of PROV-O [39, 40, 41], it is essential to formally describe the progress and interpretation of data, given different stages of processing. The same holds for applications in materials science, where Basic Formal Ontology (BFO)-based ontologies [42] describe complex data-processing workflows [43]. Such modeling strategies can be mapped to the knowledge extraction process in Viewsari to define multiple interpretive stages in the extraction and analysis of textual sources. In parallel to tracking data extraction in scientific content, which necessitates layered perspectives [20], Viewsari mirrors the evolution of interpretation across editions or commentaries in historical sources.

On equal terms, social and biographical network research shares a structural overlap. In the Sampo universe, BiographySampo [21] uses textual references to build social networks, i.e., graphs of people. Similar to how citation graphs are structured, it can be traced back to its sources; however, the approach lacks explicit provenance modeling. This concerns, for example, the different versions of a source. In the domain of music and cultural heritage, the Polifonia Ontology Network [44] exemplifies a modular approach for representing music history and the cultural heritage context in which it was generated. For example, within the module for Musical Meetups (encounters between musicians), the model supports the annotation of relationships with additional context, e.g., the time, place, and purpose [45].

In Viewsari, the goal is to build a generalizable methodology that extends these approaches with detailed, paragraph-level provenance based on PROV-O, enriched with different layers of interpretation and contextuality. Similarly, the Odeuropa model captures multifaceted information about smells [46].

The HiCO ontology (Historical Context Ontology) [47] supports provenance and interpretive assertions, providing formal representations for the differentiation of factual assertions and interpretive assertions. Viewsari draws on these approaches, making the interpretive provenance of entities and their relations explicit through linking each extracted triple to the textual fragment and the extraction process.

3. Research Questions

Three core research questions (RQs) define the scope of this Ph.D. project.

RQ1: In what way can semantic technologies model the heterogeneous content, provenance, and interpretive complexity of historical texts like Giorgio Vasari's *The Lives*?

We hypothesize that the heterogeneous and interpretive content of (historical) texts can be formalized and made machine-understandable without losing interpretive depth. This includes the development of an ontology that is capable of capturing explicit and implicit content, context, and provenance of historical sources, capturing conceptual relationships and their grounding in specific textual passages across different editions, translations, and modalities. This involves modeling provenance information for extracted entities with references to the in-line paragraphs in the different versions of the text. Thus, we assume that it is possible to represent not only the extracted entities and their relationships but also the extraction process itself to allow for traceability of provenance across different editions and interpretations.

RQ2: How can NLP methods, and LLMs in particular, be used to extract and link both explicitly and implicitly mentioned entities?

We hypothesize that the integration of LLMs into knowledge extraction pipelines enables the identification and linking of both explicitly and implicitly referenced entities in textual sources. This includes, for example, iconographic themes of artworks, motifs,[3] or vague references to their content.

Additionally, we assume that for the transformation of unstructured textual descriptions into structured representations, for explicitly mentioned entities, a mixture of state-of-the-art NLP methods and LLM prompting can increase recall and accuracy.

RQ3: What are the common challenges in modeling complex, multilingual (historical) sources, and how can a generalizable approach be drawn to make the methodology repeatable and transferable?

The goal is to map existing practices in other endeavors, e.g., from biomedicine, to the approach in Viewsari and to contribute to their methodological advancement by proposing solutions to challenges in modeling historical sources, particularly through semantic representations.

We hypothesize that a pattern-based, modular ontology engineering approach (XD + ODPs), combined with prompt-based entity extraction, can generalize to similar structured representations in other humanities and scientific domains. Using examples from Vasari's work, the goal is to identify reusable design patterns that encode recurring conceptual structures, such as multilinguality and multimodality.

4. Proposed Approach

To reflect these multidisciplinary research goals, the methodology involves different steps, i.e., knowledge extraction, ontology development, and knowledge graph construction.

In the current state of this Ph.D. project, we use the translated English edition of *The Lives* for all the knowledge extraction and engineering steps. It was written by Gaston C. Du Vere in 1912, based on the edition by Vasari released in 1568. A digital edition of Du Vere's translation is available in ten volumes, each containing a subset of all biographies [5, 48].

[3] Motifs are "recurring subject[s], theme[s], or idea[s] in art". For reference, see this site: https://blog.stephens.edu/arh101glossary/?glossary=motif.

4.1. Knowledge Extraction

For entity extraction from *The Lives*, the team of [9] experimented with different NLP models for entity recognition and entity linking, for which they created annotation guidelines [49]. The experimental pipeline plays a central role in this effort, using a range of models, including *Universal-NER* [50] for the recognition of artwork and subject recognition, and *mGENRE* [51] for the disambiguation of entities. In a corresponding GitHub repository, a sample set of the following entities is available:[4] persons, organizations, places, and miscellaneous (all following the CoNLL 2003 [52] dataset, partially linked to Wikidata[5] and Iconclass [53]), as well as artwork references, motifs, terms, and dates. For persons, the Index of Names was the baseline for extraction. As an appendix for *The Lives*, it lists all persons who occur in the work and is included in the original Italian and translated versions. In addition, co-reference resolution and statistical association from the co-occurrences of names in paragraphs model relationships between artists to create the social network.

The results are available in tabular form: CSV files summarize occurrences of entities with additional information about their provenance per volume of the translation. The data basis includes 673 co-occurrences and 1.073 persons, from which 312 appear in the relevant co-occurrences. Since this approach is only experimental, the quantity of other extracted entities is lower. For example, only 133 artworks and 311 motifs could be correctly identified [9, 54].

To expand their coverage, in ongoing work, we employ LLMs to assist in identifying and extracting further entities from the descriptions given in *The Lives*, supplementing existing annotations with OOKB entities [35, 55, 56]. A prompt-based pipeline is the methodological foundation. It integrates different LLMs, e.g., Mistral 7B [57], or LLaMA 2 13B [58], to parse text fragments such as "a painting of Madonna with Saint John in the wilderness" with dedicated prompts like "extract persons, locations, events, and artworks from the following Italian Renaissance text", and generate structured entries describing, for example, the title, motif, type, and associated artist. When not dealing with an OOKB entity, the pipeline links it to external identifiers. OOKB entities are assigned local identifiers. These candidate entities need to be manually reviewed by domain experts and stored in a separate annotation layer, as represented in the ontology. A crucial part of this pipeline is prompt engineering, for example, crafting prompts related to artistic styles or artwork characteristics based on iconographic theories can help to guide LLMs in identifying artworks or artists based on descriptive passages [13, 59].

Furthermore, another round of identifying relationships between entities is planned. This process allows the pipeline to reconstruct relationships between entities beyond just persons, such as artworks and locations, via relation extraction. To expand the scope of Viewsari, information extracted from the original Italian version of Vasari's *The Lives* will be incorporated.

4.2. Ontology Development and Knowledge Graph Construction

We develop the Viewsari ontology following the XD methodology. As an application ontology, it formally represents semi-automatically extracted content and related provenance, exemplified on the use case of *The Lives*. Using XD ensures the correct encoding of the user requirements of the art history community and provides a standardized approach to modeling and testing the conceptualizations, also based on how the data basis is shaped and what results can be achieved using the aforementioned techniques for information extraction [54, 60]. It includes both simple and complex class definitions, based on the domain needs, e.g., differentiating between co-occurrences between artists, their artwork production, or interpersonal influence. The Viewsari ontology extends different mid-level and domain ontologies to ensure interoperability [2]: For provenance representation, the ontology extends PROV-O with new classes for co-occurrences and the extraction process with LLMs. For representing the different versions of Vasari's work, Viewsari extends the Functional Requirements for Bibliographic Records (FRBR) [61] and FaBiO [62] ontologies. Subsequently, the KG construction transforms the extracted entities into a graph, using the Viewsari Ontology and Semantic Web standards such as RDF, RDFS,

[4] https://github.com/ISE-FIZKarlsruhe/vasari_nlp/

[5] https://www.wikidata.org/wiki/Wikidata:Main_Page

OWL, and SPARQL.

For the construction of the KG, we utilize Robot [63] templates and RML (RDF Mapping Language)[6] to transform the entities given in tabular form (CSV files) into RDF triples, using the ontology as a conceptual schema. For all entities, metadata, positional information (provenance), relationships, and external identifiers (e.g., Wikidata) are systematically integrated into the KG [29, 60]

All steps in ontology development and KG construction are iterative, ensuring constant external validation and evaluation in close cooperation with domain experts.

5. Preliminary Results and Evaluation

The current set of results includes: 1) a preliminary ontology to describe the complex content and contextual references of Vasari's work, distinguishing between descriptive, interpretive, and provenance-related aspects, taking into account different levels of content-description, 2) a collection of extracted entities, plus the full LLM-based pipeline for information extraction, NER, and entity linking, along with the corresponding prompts, 3) the Viewsari KG as a domain-specific prototype.

5.1. Findings from the Viewsari Pipeline

As previously mentioned, the ontology-based Viewsari KG is the core part of this ongoing Ph.D. project. The first version is based on a historical social network automatically generated as described in section 4. Another set of extracted named entities, including places, artworks, historical events, motifs, etc., complements this for additional contextualization and extension of the KG. To evaluate the first set of information extraction results, precision, recall, and F1 scores are calculated against the manually curated and annotated ground truth (as provided by [9]). The Universal-NER [50] model achieves an F1 score of 55.606% for artwork detection and 60.242% for subjects (motifs), indicating strong performance in core domains of interest, given the noisy and historical data. In the future, art history domain experts will assess the quality of LLM-generated outputs. Additionally, we aim to use the output coherence and extractability of structured data to understand which prompts achieve the highest average precision in identifying entities in art history, so they can be used and tested in other use cases. As shown in section 2, in the scientific domain, projects such as AI4DiTraRe [28] show a comparable terminological inconsistency, similarly to the hurdles in historical texts and knowledge extraction. These similarities suggest a transferable strategy for LLM-based extraction in heterogeneous corpora.

Following the XD methodology and consulting with domain experts, we could draw different conclusions about the concepts behind Viewsari: regarding the social network, a simple ontology would have sufficed to conceptualize persons and their shared relationships. However, complex relationships between entities, contextual implications of information extracted from the text, and information provenance for extracted data require a more sophisticated representation. These aspects concern all extracted resources and their relations since they originate from a specific segment of a specific work in a specific edition [2, 60]. This way, it is possible to align extracted knowledge from different versions, making them comparable on a larger scale, e.g., looking at the English translation versus the Italian original text. Accordingly, the ontology defines the formal and logical structure, covering three conceptual dimensions of the domain: bibliographic, structural, and content components. The bibliographic layer depicts information about the work and its different expressions, whereas the structural layer covers information on the document level. In the third layer, the ontology represents the extracted content (named entities, co-occurrences). As of now, the concepts in Viewsari are based on or extend ontologies such as FaBiO [62] and DoCo [64], as well as PROV-O [39] and the Web Annotation Ontology [65]. Extracted entities can be connected to source information (e.g., paragraphs in which a co-occurrence appears) from the structural layer. This then links back to the appropriate bibliographic components, allowing provenance information to be traced through detailed positional arguments and information

[6] https://rml.io/specs/rml/

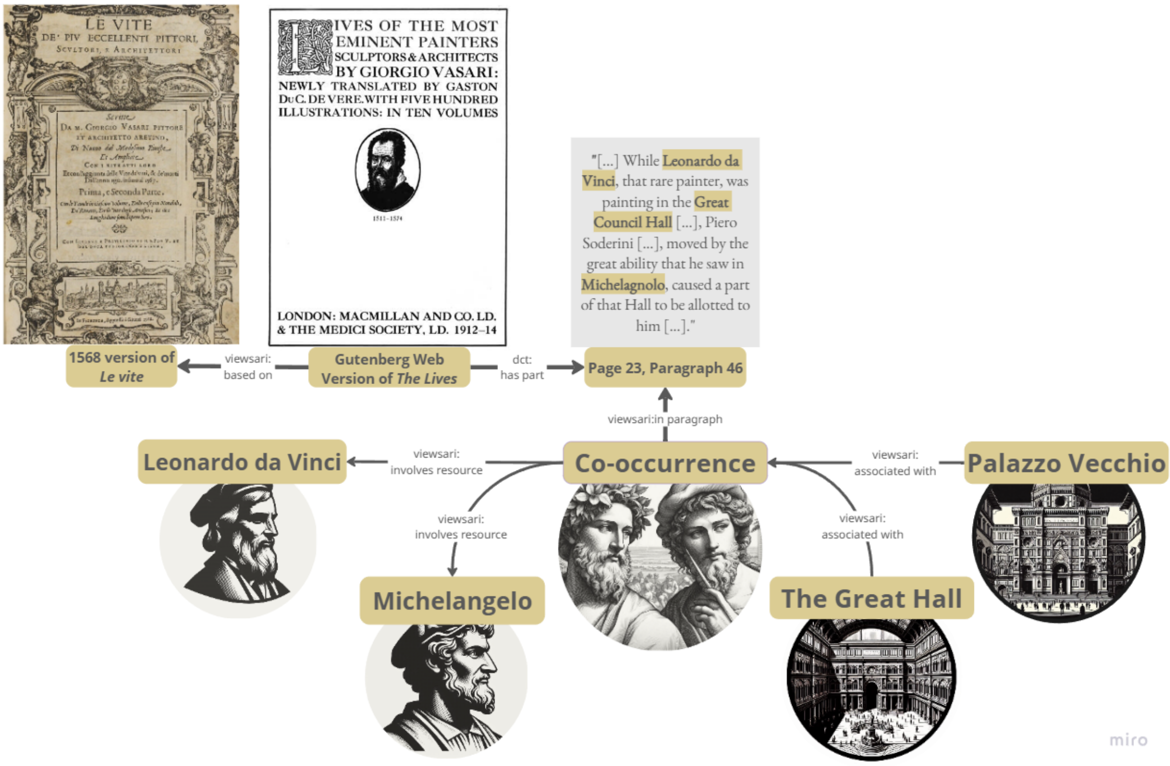
LE VITE DE' PIV ECCELLENTI PITTORI, SCVLTORI, E ARCHITETTORI
LIVES OF THE MOST EMINENT PAINTERS SCULPTORS & ARCHITECTS BY GIORGIO VASARI: NEWLY TRANSLATED BY GASTON DuC. DE VERE. WITH FIVE HUNDRED ILLUSTRATIONS: IN TEN VOLUMES
LONDON: MACMILLAN AND CO. LD. & THE MEDICI SOCIETY, LD. 1912–14
"[...] While Leonardo da Vinci, that rare painter, was painting in the Great Council Hall [...], Piero Soderini [...], moved by the great ability that he saw in Michelagnolo, caused a part of that Hall to be allotted to him [...]."
1568 version of Le vite
viewsari: based on
Gutenberg Web Version of The Lives
dct: has part
Page 23, Paragraph 46
viewsari:in paragraph
Leonardo da Vinci
viewsari: involves resource
Co-occurrence
viewsari: associated with
Palazzo Vecchio
viewsari: involves resource
viewsari: associated with
Michelangelo
The Great Hall
miro

Figure 1: The Viewsari KG: Simplified depiction of the relation of a co-occurrence with provenance information.

about the edition or translation used. Figure 1 shows a simplified visualization of the ontology.[7]

In line with the XD methodology, we evaluate all steps iteratively and user-centered. This includes validation of concepts by domain experts and prototype-based validation (based on the KG). Competency questions test the ontology's accuracy in supporting queries for domain experts [32].

5.2. Limitations

In the current version of Viewsari, the focus is solely on one corpus (the English translation by C. du Vere [48]) and thus, on one language. This limits the application of the pipeline to a narrow view in terms of knowledge extraction and transferability. For example, the generation of cross-connections between different versions can only be further highlighted when a comparison to another use case is drawn. Additionally, we currently rely almost fully on prompt-based extraction for implicitly named entities, lacking a properly benchmarked evaluation. The same holds for a lack of automated alignment or reconciliation for complex iconographic references, motifs, or entity/artwork descriptions.

5.3. Towards a Generalizable Methodology Based on Viewsari

To overcome the aforementioned limitations, one solution is to broaden the application of the methodology behind Viewsari and to consolidate various case studies. Furthermore, we address the challenges through domain expert validation for the ontology and for the results from the knowledge extraction to create a ground truth. Given the various challenges we address, we aim to make the developed solutions in Viewsari transferable to other domains. In a first attempt, as shown in prior work, we systematized reusable ODPs for DH applications [2, 60]. The goal is to further abstract and operationalize them domain-independently.

[7] For a detailed view of the ontology and alignments (T-box and A-box), please consult the Viewsari GitHub repository: https://github.com/ISE-FIZKarlsruhe/viewsari.

To assess the generalization of this methodology, we identify three key dimensions where transfer is possible: 1) Annotation modeling, namely provenance, as used in Viewsari [2]. Entities (persons, places, artworks) are modeled with links (annotations) to their textual origin. 2) Support of multiple interpretations and evolving knowledge, based on the connection where it was mentioned, who mentioned it, etc., which is also relevant to scientific knowledge graphs, where experimental results, hypotheses, and claims often require multifold descriptions. This also includes the representation of the knowledge extraction process, allowing for more detailed semantic provenance. This is true for process-oriented data, experimental parameters, and measurement provenance, as seen in biomedical data or materials science. 3) The knowledge extraction pipeline, including a variety of prompts and experiments with LLMs to see how ambiguous entities and further information can be extracted from unstructured sources.

6. Conclusion and Future Works

Integrating semantic technologies into the complex research questions of the DH community revolutionized the way knowledge can be represented, linked, and shared; however, this domain is only a singular example out of many comparable ones. However, to go one step further and improve not only the quality of the resulting ontologies but also the whole process, with this project, we aim to demonstrate how to implement a methodology to go from unstructured textual sources to formalized knowledge representations, independent of the very interdisciplinary nature of research data.

Viewsari addresses key challenges in DH and beyond, such as inconsistent modeling practices or the difficulty of representing interpretively rich, multilingual data [1, 3, 4].

Central to this project is the knowledge extraction from Vasari's text, transforming the unstructured biographies into structured resources, based on the schema provided by the developed Viewsari ontology. This enables a large-scale analysis of the corpus while retaining the provenance information of extracted entities and their relations down to the paragraph level. Additionally, the KG design bridges knowledge gaps by providing a representation level for OOKB entities, their emergence, and document-level provenance, highlighting the interpretive nature of historical data, where not all relevant knowledge has yet been curated or indexed. This also includes the recognition of entities like artworks and motifs from their description rather than by named references.

The challenges addressed in Viewsari, however, affect various domains aside from DH research. Across domains such as biomedicine, musicology, or social network research, ambiguous textual references, extracting structured knowledge with LLMs, provenance modeling, and modeling of interpretation show up alike. Thus, the goal of Viewsari is to offer a generalizable framework that can readily be adapted to other approaches.

Next steps involve broadening the Viewsari methodology by integrating more case studies. As such, we aim to support the transfer of our approach to other domains by systematizing reusable patterns and abstracting them for wider application. This also concerns a generalization of the knowledge extraction process to fit other domains, such as in scientific knowledge representation.

Acknowledgments

I would like to thank Prof. Dr. Harald Sack for his invaluable support and supervision. I also want to thank Tabea Tietz for her input and moral support during all project steps.

Declaration on Generative AI

The author has employed DeepL and Grammarly, as well as Writefull for grammar and spell checking. The author used GPT-4-turbo model for: Citation management and drafting an outline for the structure. The author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] A. Meroño-Peñuela, et al., Ontologies in CLARIAH: Towards Interoperability in History, Language and Media, CoRR abs/2004.02845 (2020). URL: https://arxiv.org/abs/2004.02845.

[2] S. R. Ondraszek, et al., One Pattern to Express Them All? Towards Generalised Patterns for Ontology Design in the Digital Humanities, in: Proceedings of the 15th Workshop on Ontology Design and Patterns (WOP 2024) at ISWC 2024, 2024. URL: https://tinyurl.com/mrxr7wnt.

[3] V. A. Carriero, et al., The Landscape of Ontology Reuse Approaches, in: Applications and Practices in Ontology Design, Extraction, and Reasoning, IOS Press, 2020. doi:10.3233/ssw200033.

[4] C. Shimizu, et al., Modular ontology modeling, Semantic Web 14 (2023) 459–489. doi:10.3233/SW-222886.

[5] G. Vasari, Le vite de' più eccellenti pittori, scultori e architettori, 2 ed., Giunti, Florence, Italy, 1568.

[6] L. Pon, Rewriting Vasari, in: The Ashgate Research Companion to Giorgio Vasari, Ashgate, 2014.

[7] S. J. Campbell, Vasari's Renaissance and Its Renaissance Alternatives, in: Renaissance Theory, Routledge, 2008.

[8] F. Moretti, Distant Reading, Verso Books, 2013.

[9] C. Santini, et al., Knowledge Extraction for Art History: the Case of Vasari's The Lives of The Artists (1568), in: Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022), 2022. doi:10.34657/10668.

[10] A. Meroño-Peñuela, et al., Semantic technologies for historical research: A survey, Semantic Web 6 (2014) 539–564. doi:10.3233/SW-140158.

[11] S. Jänicke, et al., On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges, The Eurographics Association;, 2015, pp. 83–103. doi:10.2312/eurovisstar.20151113.

[12] M. Ehrmann, et al., Named Entity Recognition and Classification in Historical Documents: A Survey, ACM Computing Surveys 56 (2023) 1–47. doi:10.1145/3604931, publisher: Association for Computing Machinery (ACM).

[13] B. Sartini, et al., ICON: An Ontology for Comprehensive Artistic Interpretations, J. Comput. Cult. Herit. 16 (2023). doi:10.1145/3594724.

[14] E. Panofsky, Studies in Iconology: Humanistic Themes in the Art of the Renaissance, Oxford University Press, New York, 1939.

[15] T. Hamaguchi, et al., Knowledge Base Completion with Out-of-Knowledge-Base Entities: A Graph Neural Network Approach, Transactions of the Japanese Society for Artificial Intelligence 33 (2018). doi:10.1527/tjsai.f-h72, publisher: Japanese Society for Artificial Intelligence.

[16] R. Peeters, et al., Entity Matching using Large Language Models, 2024. URL: https://arxiv.org/abs/2310.11244.

[17] D. G. Stork, How AI is expanding art history, Nature 623 (2023) 685–687. doi:10.1038/d41586-023-03604-3, publisher: Nature Publishing Group.

[18] A. Steiner, et al., Fine-tuning Large Language Models for Entity Matching, 2025. URL: https://arxiv.org/abs/2409.08185.

[19] E. Hyvönen, Digital humanities on the Semantic Web: Sampo model and portal series, Semantic Web 14 (2023) 729–744. doi:10.3233/SW-223034, publisher: IOS Press.

[20] T. Aggarwal, et al., Large language models for scholarly ontology generation: An extensive analysis in the engineering field, Information Processing & Management 63 (2025) 104262. doi:10.1016/j.ipm.2025.104262.

[21] E. Hyvönen, et al., BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research, in: The Semantic Web, Springer International Publishing, Cham, 2019, pp. 574–589.

[22] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticæ Investigationes 30 (2007) 3–26. doi:10.1075/li.30.1.03nad.

[23] M. Piotrowski, Natural Language Processing for Historical Texts, Springer International Publishing, Cham, 2012. doi:10.1007/978-3-031-02146-6.

[24] C. Sporleder, Natural Language Processing for Cultural Heritage Domains, Language and

Linguistics Compass 4 (2010) 750–768. URL: https://doi.org/10.1111/j.1749-818X.2010.00230.x. doi:10.1111/j.1749-818X.2010.00230.x, publisher: John Wiley & Sons, Ltd.

[25] F. De Toni, et al., Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0, in: A. Fan, et al. (Eds.), Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, Association for Computational Linguistics, virtual+Dublin, 2022, pp. 75–83. doi:10.18653/v1/2022.bigscience-1.7.

[26] A. Graciotti, Knowledge Extraction from Multilingual and Historical Texts for Advanced Question Answering, in: Proceedings of ISWC 2023, ISWC 2023 Doctoral Consortium, CEUR Workshop Proceedings (CEUR-WS.org), Athens, Greece, 2023.

[27] A. M. Brasoveanu, et al., In Media Res: A Corpus for Evaluating Named Entity Linking with Creative Works, in: R. Fernández, T. Linzen (Eds.), Proceedings of the 24th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2020, pp. 355–364. doi:10.18653/v1/2020.conll-1.28.

[28] A. Jacyszyn, et al., AI4DiTraRe: Towards LLM-Based Information Extraction for Standardising Climate Research Repositories, in: First AAAI Bridge on Artificial Intelligence for Scholarly Communication AI4SC, Philadelphia, United States, 2025. doi:10.5281/zenodo.14872358.

[29] S. Barzaghi, et al., CHAD-KG: A Knowledge Graph for Representing Cultural Heritage Objects and Digitisation Paradata, 2025. URL: https://arxiv.org/abs/2505.13276.

[30] A. Gangemi, V. Presutti, Ontology Design Patterns, in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 221–243. doi:10.1007/978-3-540-92673-3_10.

[31] V. Presutti, et al., eXtreme design with content ontology design patterns, in: Proceedings of the 2009 International Conference on Ontology Patterns - Volume 516, WOP'09, CEUR-WS.org, Aachen, DEU, 2009, pp. 83–97. Event-place: Washington DC.

[32] E. Blomqvist, et al., Engineering Ontologies with Patterns - The eXtreme Design Methodology., in: Ontology Engineering with Ontology Design Patterns - Foundations and Applications, 2016, pp. 23–50. doi:10.3233/978-1-61499-676-7-23.

[33] V. Carriero, et al., ArCo: The Italian Cultural Heritage Knowledge Graph, 2019, pp. 36–52. doi:10.1007/978-3-030-30796-7_3.

[34] Y. Tzitzikas, et al., CIDOC-CRM and Machine Learning: A Survey and Future Research, Heritage 5 (2022) 1612–1636. doi:10.3390/heritage5030084.

[35] A. Ahola, L. Peura, H. Rantala, Using generative AI and LLMs to enrich art collection metadata for searching, browsing, and studying art history in Digital Humanities, 2024.

[36] N. Ockeloen, et al., BiographyNet: managing provenance at multiple levels and from different perspectives, in: Proceedings of the 3rd International Conference on Linked Science - Volume 1116, LISC'13, CEUR-WS.org, 2013, p. 59–71.

[37] M. Kienle, Between Nodes and Edges: Possibilities and Limits of Network Analysis in Art History, Artl@s Bulletin 6 (2017).

[38] P. Ciccarese, et al., PAV ontology: provenance, authoring and versioning, Journal of Biomedical Semantics 4 (2013) 37. doi:10.1186/2041-1480-4-37.

[39] T. Lebo, et al., PROV-O: The PROV ontology, Technical Report, World Wide Web Consortium, 2013.

[40] S. S. Sahoo, et al., Scientific Reproducibility in Biomedical Research: Provenance Metadata Ontology for Semantic Annotation of Study Description, in: AMIA Annual Symposium Proceedings, volume 2016, American Medical Informatics Association, 2016, pp. 1070–1079.

[41] T. Procko, O. Ochoa, Mapping the W3C Provenance Ontology (PROV-O) to the Basic Formal Ontology (BFO): Epistemological Considerations and Preliminary Implementation, Social Science Research Network (2024). URL: http://dx.doi.org/10.2139/ssrn.4852748.

[42] B. Smith, et al., Basic Formal Ontology (BFO) Version 2020-08-26, 2020. URL: http://purl.obolibrary.org/obo/bfo.owl.

[43] B. Bayerlein, et al., PMD Core Ontology: Achieving semantic interoperability in materials science, Materials & Design 237 (2024) 112603. doi:10.1016/j.matdes.2023.112603.

[44] J. de Berardinis, et al., The Polifonia Ontology Network: Building a Semantic Backbone for Musical Heritage, in: The Semantic Web – ISWC 2023, Springer Nature Switzerland, Cham, 2023, pp. 302–322.

[45] A. M. Tirado, et al., Musical Meetups: a Knowledge Graph approach for Historical Social Network Analysis, in: ESWC 2023 Workshops and Tutorials. Semantic Methods for Events and Stories (SEMMES), CEUR Workshop Proceedings (CEUR-WS.org), 2023. URL: https://oro.open.ac.uk/88720/.

[46] P. Lisena, et al., Capturing the Semantics of Smell: The Odeuropa Data Model for Olfactory Heritage Information, in: The Semantic Web, Springer International Publishing, Cham, 2022, pp. 387–405. doi:10.1007/978-3-031-06981-9_23.

[47] M. Daquino, F. Tomasi, Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects, in: E. Garoufallou, R. J. Hartley, P. Gaitanou (Eds.), Metadata and Semantics Research, Springer International Publishing, Cham, 2015, pp. 424–436. doi:https://doi.org/10.1007/978-3-319-24129-6_37.

[48] G. Vasari, Lives of the Most Eminent Painters, Sculptors and Architects, Project Gutenberg, 2008. URL: https://onlinebooks.library.upenn.edu/webbin/gutbook/lookup?num=25326.

[49] C. Santini, et al., Guidelines for the Annotation of ExtrART: Evaluation Dataset for Entity Extraction from The Lives Of The Artists (1550), 2023. doi:10.5281/zenodo.7991340.

[50] W. Zhou, et al., UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition (2023).

[51] N. De Cao, et al., Multilingual Autoregressive Entity Linking, Transactions of the Association for Computational Linguistics 10 (2022) 274–290. doi:10.1162/tacl_a_00460.

[52] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142–147. URL: https://www.aclweb.org/anthology/W03-0419.

[53] H. Brandhorst, E. Posthumus, Iconclass: a key to collaboration in the digital humanities, in: The Routledge Companion to Medieval Iconography, Routledge, 2016, pp. 201–218.

[54] S. R. Ondraszek, et al., Viewsari: New Perspectives on Historical Network Analysis in Giorgio Vasari's The Lives Using Knowledge Graphs, in: Historical Network Research, Zenodo, Lausanne, 2024. doi:10.5281/zenodo.1260671.

[55] T. Sztyler, et al., Lode: Linking digital humanities content to the web of data, in: IEEE/ACM Joint Conference on Digital Libraries, 2014, pp. 423–424. doi:10.1109/JCDL.2014.6970206.

[56] J. Baas, et al., Entity Matching in Digital Humanities Knowledge Graphs, 2021, pp. 1–15. URL: https://2021.computational-humanities-research.org/.

[57] A. Q. Jiang, et al., Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825.

[58] H. Touvron, et al., Llama 2: Open foundation and fine-tuned chat models, 2023. URL: https://arxiv.org/abs/2307.09288.

[59] B. Sartini, IICONGRAPH: Improved iconographic and iconological statements in knowledge graphs, in: The Semantic Web: 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26–30, 2024, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2024, p. 57–74. doi:10.1007/978-3-031-60635-9_4.

[60] S. R. Ondraszek, et al., eXtreme Design for Ontological Engineering in the Digital Humanities with Viewsari, a Knowledge Graph of Giorgio Vasari's The Lives, in: Proceedings of the 1st International Workshop on Semantic Digital Humanities (SemDH 2024), CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3724/paper5.pdf.

[61] B. Tillett, What is FRBR? A conceptual model for the bibliographic universe, The Australian Library Journal 54 (2005) 24–30. doi:10.1080/00049670.2005.10721710.

[62] S. Peroni, D. Shotton, FaBiO and CiTO: Ontologies for describing bibliographic resources and citations, Journal of Web Semantics 17 (2012) 33–43. doi:10.1016/j.websem.2012.08.001.

[63] R. C. Jackson, et al., ROBOT: A Tool for Automating Ontology Workflows, BMC Bioinformatics 20 (2019) 407. doi:10.1186/s12859-019-3002-3.

[64] A. Constantin, et al., The DocumentComponents Ontology(DoCO), Semantic Web 7 (2016) 167–181. doi:10.3233/SW-150177.

[65] P. Ciccarese, et al., An open annotation ontology for science on web 3.0, Journal of Biomedical Semantics 2 (2011) S4. doi:10.1186/2041-1480-2-S2-S4.