ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

PATRIMONIO CULTURALE NELL'ECOSISTEMA DIGITALE

Ciclo 38

Settore Concorsuale: 01/B1 - INFORMATICA

Settore Scientifico Disciplinare: INF/01 - INFORMATICA

STRUCTURING CULTURAL HERITAGE CONTENT AND CONTEXT: INTEGRATING LLMS IN ONTOLOGY-DRIVEN KNOWLEDGE GRAPH EXTRACTION

Presentata da: Andrea Schimmenti

Coordinatore Dottorato

Francesca Tomasi

Supervisore

Fabio Vitali

Co-supervisore

Maria Godefrida Jacoba Van Erp

Esame finale anno 2026

I, Andrea Schimmenti, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Cultural Heritage institutions have digitized extensive collections and published their metadata through Semantic Web technologies, yet the content and the scholarly contextualization of documents—entities, relationships, events, and interpretations—remains largely inaccessible through semantic querying. Manual Knowledge Graph creation proves prohibitively expensive at scale, while automatic Knowledge Extraction faces critical barriers in CH contexts: limited annotated training data and domain-specific linguistic complexity. This dissertation investigates automatic Knowledge Graph extraction from Cultural Heritage texts in data-scarce scenarios, addressing three research questions: (1) What methodologies and challenges characterize existing CH text-to-KG projects? (2) How can Large Language Models be integrated into ontology-driven Knowledge Extraction pipelines, and what are the limitations and trade-offs? (3) Can LLM-based systems produce sufficiently accurate Knowledge Graphs of scholarly interpretations while preserving provenance and epistemic uncertainty? We conduct a systematic survey of eleven CH projects (2015-2025) and analyze 227 papers, identifying persistent bottlenecks in Named Entity Recognition, Relationship Extraction, and Entity Linking. We introduce *Adaptive Text-to-KG for Cultural Heritage* (ATR4CH), a five-step methodology coordinating ontology analysis, Competency Question formulation, ground-truth annotation, LLM-based extraction, and multi-layered evaluation. We validate ATR4CH through case studies including authenticity debates, archival finding aids, RAG-based argument extraction, and synthetic training data generation for Aspect-Based Sentiment Analysis. Results

establish that LLMs enable ontology-aligned extraction under data scarcity, achieving accuracy sufficient for scholarly workflows. LLMs augment rather than replace traditional pipelines, providing capabilities for bootstrapping development and serving domains where annotation costs cannot be justified. However, human oversight remains necessary: errors may propagate through pipelines, data alignment represents a persistent bottleneck, and epistemic uncertainty requires continued development. This dissertation advances the state of the art by providing a replicable methodological framework and empirical evidence that LLM-based extraction can bridge the gap between digitization and semantic accessibility of Cultural Heritage repositories.

Table of Contents

Abstract **i**

Table of Contents **iii**

Abbreviations **ix**

List of Figures **xi**

List of Tables **xiv**

List of Code Listings **xvi**

Introduction **1**

Research Questions 7

Outline 8

1 Background **11**

1.1 Knowledge Representation for Cultural Heritage 12

1.2 Knowledge Extraction Paradigms 16

1.2.1 Closed Knowledge Extraction 17

1.2.2 Open Knowledge Extraction 19

1.2.3 Fine-Grained Approaches 20

1.2.4 Knowledge Extraction for the Semantic Web 22

1.3 Chapter Summary 23

2 Text-to-KG in the Cultural Heritage domain 25

2.1 Background . . . 26

2.2 Methodology . . . 27

2.2.1 Project Identification Strategy . . . 27

2.2.2 Literature Analysis for Trend Identification . . . 29

2.3 Case Studies . . . 30

2.3.1 Structured and Semi-Structured Sources . . . 31

2.3.2 Unstructured Narrative Sources . . . 35

2.3.3 Specialized Documents . . . 43

2.3.4 Comparative Analysis Across Projects . . . 46

2.4 Challenges . . . 49

2.4.1 Data Alignment Challenges . . . 49

2.4.2 Language Processing Challenges . . . 52

2.4.3 Entity Recognition and Specialized Vocabularies . . . 53

2.4.4 Document Quality and Evaluation . . . 54

2.5 Trends . . . 55

2.5.1 Named Entity Recognition for Specialized Contexts . . . 56

2.5.2 Relation Extraction Advances . . . 57

2.5.3 Addressing the Long-Tail Entity Linking Challenge . . . 59

2.5.4 Evaluation Methodologies Beyond Traditional Metrics . . . 60

2.5.5 LLMs in Ontology Engineering and Text-to-KG . . . 62

2.6 Defining Text-to-KG for Cultural Heritage . . . 64

2.6.1 The Dominant Pipeline Architecture . . . 67

2.7 Conclusion . . . 70

3 Towards LLM-assisted pipelines for text-to-KG 75

3.1 LLM-assisted text-to-KG modules . . . 76

3.2 LLMs and Implicit Information . . . 79

3.3 LLMs within text-to-KG: Defining scope and boundaries . . . 82

3.3.1 Advantages . . . 82

3.3.2 Limitations . . . 82

3.3.3 Hybrid architectures as pragmatic solutions 83
3.3.4 Open problems for LLM integration 84
3.4 Conclusions 86

4 Knowledge Extraction on Archival Finding Aids 87
4.1 Background 88
4.2 The Records in Contexts Ontology 90
4.3 Methodology 91
4.3.1 Evaluation Framework 91
4.4 Pipeline Implementation and Case Study 93
4.4.1 Step 1: Text Classification 96
4.4.2 Step 2: Question Answering 97
4.4.3 Step 3: Schema-Based JSON Generation 98
4.4.4 Post-processing and RDF Mapping 100
4.5 Results 100
4.5.1 Structural Level 100
4.5.2 Information Level 103
4.5.3 Interpretative Level 104
4.6 Discussion 105
4.7 Conclusion 111

5 Knowledge Extraction in the Curation life cycle: The ATR4CH Methodology 114
5.1 Background 115
5.2 Theoretical Foundations 117
5.3 The ATR4CH Methodology 119
5.3.1 Foundational Analysis and Design (Task I) 125
5.3.2 Minimal Working Annotation Development (Task II) . . 126
5.3.3 Pipeline Architecture Development (Task III) 129
5.3.4 Integration and Refinement (Task IV) 130
5.3.5 Knowledge Extraction and Evaluation (Task V) 131

5.4 Conclusion 133

6 Ontology-driven Opinion Mining in Authenticity Assessment Debates 135

6.1 Problem Statement 138
6.2 Research Questions 138
6.3 Outline 139
6.4 Related Work 140
6.4.1 Knowledge Representation 140
6.4.2 Opinion Mining for the Semantic Web 142
6.5 Methodology and Materials 142
6.5.1 Foundational Inputs (Task I) 142
6.5.2 Minimal Working Annotation Development (Task II) . . 150
6.5.3 Production Annotation Model (Task IV) 153
6.6 Knowledge Extraction Pipeline and Evaluation Framework . . . 158
6.6.1 Sequential Processing Pipeline 160
6.6.2 Knowledge Graph Generation 166
6.6.3 Evaluation Framework 169
6.7 Results 173
6.7.1 EQ1: CH Item Metadata Extraction Precision 174
6.7.2 EQ2: Scholarly Entity Recognition Coverage 175
6.7.3 EQ3: Evidential Reasoning Extraction Quality 176
6.7.4 EQ4: Hypothesis and Judgment Identification 177
6.7.5 EQ5: Overall Discourse Representation Fidelity 177
6.8 Discussion and Conclusions 178
6.8.1 Extraction Performance Analysis 178
6.8.2 Representation Fidelity and Quality Assessment 179
6.8.3 Model Comparison and Performance Trade-offs 179
6.8.4 Methodological Framework Validation 180
6.8.5 Deployment Implications and Cost-Effectiveness 181
6.8.6 Contributions, Limitations and Future Directions 181

7 Knowledge Extraction of the *Van den Vos Reynaerde* Authorship Debate 183

7.1 Background . . . 185

7.1.1 Willem die Madocke maecte . . . 185

7.1.2 The Digital Hermeneutics Model . . . 186

7.1.3 Beyond Paragraph-level Text-to-KG with RAG . . . 189

7.2 Methodology . . . 190

7.2.1 Corpus Selection and Research Questions . . . 191

7.2.2 Ontological Model for Authorial Attribution . . . 193

7.2.3 Pipeline Architecture and Implementation . . . 199

7.3 Results . . . 204

7.3.1 Authorship Attribution Interpretations . . . 205

7.4 Conclusions and Future Work . . . 215

8 Synthetic Training Data Generation for Aspect-Based Sentiment Analysis on Book Reviews 219

8.1 Related Work . . . 222

8.1.1 Aspect-Based Sentiment Analysis . . . 222

8.1.2 Synthetic Dataset Generation . . . 223

8.1.3 Entity Typing . . . 224

8.1.4 LLMs for Knowledge Extraction . . . 224

8.2 Data and Resources . . . 224

8.2.1 Book Reviews Dataset . . . 225

8.2.2 Structured Data . . . 225

8.2.3 DOLCE . . . 225

8.2.4 Llama 3.1 . . . 226

8.3 Methodology . . . 226

8.3.1 Semi-synthetic Dataset Generation . . . 226

8.3.2 Model Adaptation . . . 230

8.4 Evaluation . . . 231

8.4.1 Llama 3.1-8B Instruct Baseline . . . 231

8.4.2 Llama 3.1-8B ABSA+ET Fine-Tuned Model 232
8.5 Conclusions and Future Work 236

Conclusions **240**
Discussion 248
Final Remarks 253

Appendices **260**
Survey Queries 260
Survey Tables 263

Bibliography **274**

List of Abbreviations

Concept	Abbreviation
Adaptive Text-to-KG for Cultural Heritage	ATR4CH
Aspect-based Sentiment Analysis	ABSA
Abstract Meaning Representation	AMR
Basic Formal Ontology	BFO
Competency Question	CQ
Cultural Heritage	CH
CIDOC Conceptual Reference Model	CIDOC CRM
CRM Inference	CRMInf
Digital Humanities	DH
Descriptive Ontology for Linguistic and Cognitive Engineering	DOLCE
Dublin Core Metadata Initiative	DCMI
Entity Linking	EL
Evaluation Question	EQ
Findable, Accessible, Interoperable, Reusable	FAIR
Libraries, Archives, Museums	LAM
Ground Truth	GT
Historical Context Ontology	HiCo
Information Extraction	IE
Knowledge Base	KB
Knowledge Extraction	KE

Concept	**Abbreviation**
Knowledge Graph	KG
Large Language Model	LLM
Linked Open Data	LOD
Low-Rank Adaptation	LoRA
Named Entity Recognition	NER
Natural Language Processing	NLP
Optical Character Recognition	OCR
Pre-trained Language Model	PTLM
Provenance Ontology	PROV-O
Quantized Low-Rank Adaptation	QLoRA
Retrieval-Augmented Generation	RAG
Records in Contexts Ontology	RiC-O
Resource Description Framework	RDF
Relation Extraction	RE
Research Question	RQ
Scholarly Evidence-Based Interpretation	SEBI
Sentiment Analysis	SA
Scholarly Digital Edition	SDE
Semantic Web	SW
SPARQL Protocol and RDF Query Language	SPARQL
Text Encoding Initiative	TEI
Text-to-Knowledge Graph	Text-to-KG
Web Ontology Language	OWL

List of Figures

2.1 Flowchart of the modular text-to-KG pipeline architecture observed across surveyed CH projects 67

4.1 RIC-O Event core classes and properties 94

4.2 RIC-O Relation core classes and properties 94

4.3 The birth of Andrea Costa 101

4.4 the relationship between Andrea Costa and Anna Kuliscioff . . . 101

4.5 Output evaluation web application. In the example, the Birth Event of Andrea Costa is being evaluated. 102

5.1 Workflow of the ATR4CH methodology showing the iterative task structure. 123

6.1 The Donation of Constantine entry in Wikidata 137

6.2 Overall distribution of article lengths showing the right-skewed pattern characteristic of encyclopedic content, with most articles in the 2k–15k character range and notable outliers extending beyond 40k characters 144

6.3 Distribution of articles across Wikipedia categories, showing the natural prevalence of different forgery types in scholarly discourse 145

6.4 Token count distribution by category, illustrating variability in article length and content density. Box plots show medians, quartiles, and outliers representing comprehensive case studies . 146

6.5 Plato noted as the author of the Demodocus using a deprecated rank, illustrating how existing knowledge bases can represent disputed attributions 146

6.6 Selection of classes and properties to represent scholarly claims addressing authenticity assessment of a document 148

6.7 Selection of classes and properties to represent contextual information about scholarly claims addressing authenticity assessment of a document 149

6.8 Example annotation of an entity expressing an opinion about a CH item 151

6.9 Alleged metadata annotation for the Donation of Constantine . 153

6.10 Lorenzo Valla's opinion with feature assessment annotation . . . 154

6.11 Caesar Baronius's admission of forgery with provenance annotation 156

6.12 Johannes Fried's hypotheses annotation for the Donation of Constantine 156

6.13 Flowchart of the sequential pipeline for SEBI-based KG generation 159

6.14 CH Item metadata extraction from source text to structured JSON output 161

6.15 Cognizer identification from source text to structured JSON output 162

6.16 Entity resolution and linking: from clustered mentions to Wikidata-enriched entities 163

6.17 JSON output from source text about Cognizer, subject of the opinion, provenance, and assessment 164

6.18 JSON output of identified evidence with evaluation 165

6.19 Hypothesis extraction from source text to structured alternative theories 166

6.20 Lorenzo Valla's statement about the Donation of Constantine . 167

6.21 SPARQL query used to extract entity counts from the KGs for statistical comparison across models 173
6.22 Radar Chart of different KG extractions 174

7.1 Example of a nanopublication about malaria. Source: Groth et al., *The Anatomy of a Nano-publication* [82] 187
7.2 Factual layer representation (Graffoo diagram) 194
7.3 Model to represent the *Van den vos Reynaerde* creation event (assertion layer, Graffoo diagram) 195
7.4 Model to represent the Reynaertdichter (assertion layer, Graffoo diagram) 196
7.5 Layer 2 - Interpretation layer (Graffoo diagram) 197
7.6 Layer 3 - Provenance layer (Graffoo diagram) 198

8.1 Synthetic Dataset Generation Pipeline for ABSA-Annotated Book Reviews 226
8.2 JSON schema for aspect extraction 230

List of Tables

2.1 Dimensions of Analysis for Text-to-KG Approaches 29

2.2 Overview of Text-to-KG Projects in Cultural Heritage 31

2.3 Comparative Analysis of Text-to-KG Projects 46

3.1 Performance comparison of LLMs on implicit versus explicit information extraction tasks. Results demonstrate model size effects on handling implicit reasoning. 81

4.1 Valid outputs percentage 103

4.2 Macro-average metrics 104

4.3 Micro-average precision, recall, and F_1 score per class 104

4.4 Qualitative evaluation results per event class 106

6.1 Distribution of articles across Wikipedia categories in the corpus 144

6.2 Ground Truth annotation results 158

6.3 KE overall metrics 174

6.4 CH Item Metadata extraction performance across three LLMs . 175

6.5 Entity recognition coverage and accuracy 176

6.6 Evidence extraction quality and coverage 176

6.7 Hypothesis and judgment extraction performance 177

6.8 Per-statement Correctness (G-EVAL scores on 0-1 scale) 178

6.9 Overall Debate Representativeness (G-EVAL scores on 0-1 scale) 178

7.1 Mapping from schema relations to CIDOC CRM patterns 204

7.2 Extraction statistics for nanopublications generated from scholarly articles (temperature 0.3) 205
7.3 Competency question coverage across extracted KGs 211

8.1 DOLCE type distribution in the dataset 228
8.2 Llama 3.1-8B Instruct Performance Metrics. The Predicted Aspects percentage (200.11%) indicates that the model generated approximately twice as many aspects as exist in the ground truth 232
8.3 Llama 3.1-8B ABSA+ET Performance Metrics. The Predicted Aspects percentage (131.30%) indicates that the fine-tuned model generated about 31% more aspects than in the GT, showing improved precision compared to the baseline Instruct model 233
8.4 Fine-Tuned Model Performance Summary 233
8.5 Distribution Comparison Between Model and Ground Truth . . 234
8.6 BiographySampo: Text-to-KG Implementation 263
8.7 ParliamentSampo: Parliamentary Data as KGs 264
8.8 WarSampo: World War II Heritage as Linked Data 265
8.9 HyperReal: Cultural Symbolism KG 266
8.10 InTaVia: Transnational Biographical Heritage Integration 267
8.11 Odeuropa: Olfactory Heritage Knowledge Extraction 268
8.12 Notarypedia: Legal Heritage Knowledge Extraction 269
8.13 MMKG: Musical Meetups KG 270
8.14 MusicBO: Bologna Musical Heritage KG 271
8.15 Viewsari: Renaissance Art History KG 272
8.16 Bernoulli-Euler Digital: Historical Scientific Correspondence Knowledge Extraction 273

List of Code Listings

4.1 First prompt for text classification 96
4.2 Second prompt for question answering 97
4.3 Event schema for the class "POLITICS" 99
4.4 JSON output of the model from the text (simplified) 106
4.5 Part of the RDF version of the JSON in Listing 4.4 107
6.1 Core Annotation Mapping Algorithm 151
6.2 Evidence and Feature Mapping Algorithm 154
6.3 Hypotheses Mapping Algorithm 156
6.4 Document representation with alleged and scholarly metadata . 167
6.5 Lorenzo Valla's interpretation with supporting evidence 168
6.6 Lorenzo Valla's philological evidence structure 169
7.1 Namespace prefixes for the ontological model 193
7.2 Assertion layer of Van Daele (2005) 206
7.3 Assertion layer of Besamusca & Bouwman (2009) 207
7.4 Assertion layer of Peeters (1973) 209
7.5 Assertion layer of Wackers (2016) 209

Introduction

The problems are solved, not by giving new information,
but by arranging what we have known since long.

— Wittgenstein, Philosophical Investigations (1953)

The objective of this dissertation is to investigate how Large Language Models can enable the automatic extraction of Knowledge Graphs from Cultural Heritage texts, particularly in scenarios where limited annotated training data exists but established ontological models are available to guide knowledge representation. This work addresses a structural tension in digital Cultural Heritage: while decades of digitization efforts have successfully brought extensive collections into digital repositories and published their metadata through Semantic Web technologies, the intellectual content within these documents remains largely inaccessible through semantic querying [78]. Consider a manuscript digitized and cataloged according to established standards: its metadata (creator, date, provenance, physical characteristics) can be queried across institutional boundaries through standardized ontologies and Linked Open Data technologies. Yet, in most cases, the entities mentioned within its text, the relationships between those entities, the events described, and the scholarly interpretations embedded in commentary remain in unstructured prose, accessible only through basic string matching [147, 25]. This represents a significant limitation. Scholars studying patterns across collections, institutions seeking to enable sophisticated discovery interfaces, and researchers attempting to trace intellectual networks often must resort to manual reading or keyword searches that treat carefully crafted historical narratives as bags

of words. The challenge is not only technical but economic and methodological: generating structured knowledge graphs from textual content at scale requires either prohibitive manual effort or automatic extraction approaches that can operate with limited training data while maintaining the precision demanded by scholarly applications. Manual annotation by domain experts provides high accuracy but scales poorly to collections containing thousands or millions of documents. Traditional supervised learning approaches require extensive annotated corpora, resources that remain unavailable for most specialized Cultural Heritage domains and historical languages. The scalability bottleneck persists despite successful demonstrations of structured content representation in smaller collections. Before examining this challenge in detail, three key concepts require definition. **Knowledge Graphs** (KGs), as understood within the Semantic Web framework, are structured representations of information in the form of subject-predicate-object triples encoded using the Resource Description Framework (RDF),[1] where each element is identified by Uniform Resource Identifiers (URIs) that enable unambiguous information sharing across systems [22, 25]. The structure and semantics of these graphs are defined by formal ontologies expressed in the Web Ontology Language (OWL),[2] which specifies classes, properties, and the relationships between them [201]. Knowledge graphs can be queried using SPARQL, a standardized query language for RDF data.[3] **Cultural Heritage** (CH) in this work encompasses materials held by libraries, archives, and museums: manuscripts, historical documents, scholarly texts, archival finding aids, and born-digital cultural artifacts. **Large Language Models** (LLMs) are transformer-based neural architectures trained on extensive text corpora, demonstrating capabilities for few-shot learning (the ability to perform tasks with minimal training examples) and semantic understanding of complex discourse structures [27]. The digitization efforts of recent decades have transformed paradigms of cul-

[1] `https://www.w3.org/RDF/`
[2] `https://www.w3.org/OWL/`
[3] `https://www.w3.org/TR/sparql11-query/`

tural preservation and access. Cultural institutions have invested substantial resources in systematic digitization that has created extensive repositories of digital objects [193]. These efforts have been guided by well-established frameworks: the DCC Curation Lifecycle Model[4] provides systematic guidance for preservation and management of digital materials, focusing on technical preservation aspects and metadata management throughout conceptualization, creation, evaluation, ingest, storage, access, and transformation phases [92, 199]. Many institutions have implemented this lifecycle using Semantic Web technologies to publish collection metadata, enriching the Linked Open Data cloud [201]. Knowledge graphs built using RDF provide the technical foundation, storing information as triples that can be queried through SPARQL. Ontologies provide the semantic layer: domain-specific models such as CIDOC CRM[5] for cultural heritage objects, Dublin Core[6] for bibliographic materials, and PROV-O[7] for provenance documentation define standardized classes and properties [25, 201]. Controlled vocabularies such as the Getty Vocabularies[8] support precise semantic annotation across domains [88]. These technologies have enabled sophisticated cross-institutional discovery that was previously impossible. Europeana[9] exemplifies this achievement: a platform aggregating digital collections from thousands of European institutions, employing Semantic Web technologies to enable unified querying across heterogeneous sources. The Digital Humanities community has built extensive infrastructure, tools, and applications on this foundation [193, 201]. Yet this success story remains incomplete. The large-scale digitization process has focused predominantly on publishing metadata (descriptions of what resources are, when they were created, and who was involved in their creation). Such metadata is essential for resource identification and provenance preservation in the Linked Open Data cloud. However, the content of the resources themselves has not been system-

[4] `https://www.dcc.ac.uk/guidance/curation-lifecycle-model`
[5] `http://www.cidoc-crm.org/`
[6] `https://dublincore.org/`
[7] `https://www.w3.org/TR/prov-o/`
[8] `http://www.getty.edu/research/tools/vocabularies/`
[9] `https://www.europeana.eu/`

atically included in this process [78]. For textual resources, information about which entities are mentioned throughout a text, which relationships exist between those entities, which events are described, and which scholarly interpretations are embedded in the discourse remains largely absent from published knowledge bases. The consequences extend beyond simple inconvenience. Current information retrieval systems in most cultural heritage institutions leave rich textual content largely inaccessible through semantic querying. Discovering connections that span collections requires manual reading rather than computational query. Patterns that could support both qualitative scholarly analysis and quantitative studies remain hidden [43, 150, 147]. Knowledge graphs containing structured content information could serve multiple functions: as sources of curated knowledge for scholarly analysis, as demonstrated by academic community usage of Wikidata [228];[10] as foundations for intelligent question-answering systems serving researchers and the public [223]; and as infrastructures enabling new forms of scholarly inquiry. A pioneering project demonstrates this potential. *Vespasiano da Bisticci, Lettere. Knowledge Site 3.0*[11] represents the epistolary corpus of a fifteenth-century Italian humanist and book merchant [200]. Beyond encoding letter metadata and text in XML-TEI,[12] the project described each entity appearing in the letters through knowledge graphs linked to authority files. This relatively modest addition—describing persons, places, and codices mentioned in the correspondence—enables multi-dimensional querying not only by correspondents or dates but also by mentioned entities, referenced locations, and manuscript names. The resulting interface supports scholarly investigations that would be impossible with metadata alone [200]. However, this project encompasses only forty-five letters. The manual effort required to produce such structured representations does not scale to larger collections. Domain experts must not only identify

[10] Wikidata (`https://www.wikidata.org/`) is a collaboratively edited, multilingual knowledge base maintained by the Wikimedia Foundation, containing over 100 million structured data items.

[11] `https://projects.dharc.unibo.it/vespasiano/`

[12] `https://tei-c.org/`

named entities, but also recognize implicit relationships, temporal sequences, causal connections, and contextual dependencies, often expressed through subtle linguistic patterns rather than explicit statements. Experts require training in annotation practices, and quality assurance through evaluation remains necessary. For collections containing thousands or millions of documents, structured manual annotation is confronted with an economic barrier. This scalability bottleneck has motivated research in automatic Knowledge Extraction, the field that addresses subtasks necessary to produce knowledge graphs from textual input [132, 95]. These approaches typically orchestrate multiple Natural Language Processing tasks: Named Entity Recognition identifies and classifies entity mentions, Relation Extraction determines semantic connections between entities, and Entity Linking maps mentions to canonical identifiers in reference knowledge bases [95]. Berners-Lee's original vision for the Semantic Web anticipated this challenge, envisioning intelligent software agents capable of harvesting and processing information across distributed sources: "The real power of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs" [22]. Extending this vision to the rich textual content of cultural collections would enable scholars to query not merely what documents exist but what ideas, relationships, events, and interpretations are expressed within them. The Odeuropa project demonstrates large-scale automatic Knowledge Extraction in a cultural heritage context [120]. This Horizon 2020 initiative processed 167,029 textual resources and 43,679 visual materials to construct a European Olfactory Knowledge Graph containing 2.4 million smell instances across six languages. Before structured extraction, researchers could only search for literal mentions of smell-related terms. The extracted knowledge graphs enabled sophisticated querying through dedicated navigation paths organized by smell sources, fragrant spaces, and olfactory gestures [122]. However, even this successful project faced domain-specific challenges requiring substantial methodological adaptation. Modern Natural Language

Processing frameworks provide robust capabilities for fundamental text processing. Libraries such as spaCy offer industrial-strength processing with pretrained models for Named Entity Recognition, dependency parsing, and part-of-speech tagging across multiple languages [94]. Flair provides state-of-the-art contextual embeddings and sequence labeling [5]. Yet applying these general-purpose tools to Cultural Heritage documents presents persistent difficulties. Historical texts contain specialized vocabularies, temporal language variation, and domain-specific entity types underrepresented in standard training corpora [60]. Adapting these models requires techniques beyond basic transfer learning, and performance degradation remains substantial without domain-specific training data. The development of transformer-based architectures and Large Language Models has introduced capabilities that may address data scarcity challenges. These models demonstrate few-shot learning (performing tasks with minimal training examples) and domain adaptation without extensive labeled datasets [27]. In contexts where annotated training data are limited or unavailable, conditions common in Cultural Heritage domains, LLMs show potential for both automatic data annotation and direct structured extraction [113, 170, 78]. Their ability to process and generate structured data according to specified schemas makes them potentially valuable for extending Semantic Web paradigms from metadata to content analysis. However, responsible implementation demands careful attention to limitations. Cultural Heritage applications require high precision to maintain academic credibility, necessitating rigorous evaluation of automated extraction accuracy, proper provenance tracking of machine-generated annotations, and integration with knowledge organization systems developed through decades of careful curation [199]. The cost of false positives (incorrectly identified relationships or misattributed entities) can undermine trust in digital scholarly resources. False negatives represent missed opportunities to surface valuable connections that could support research discoveries. Understanding accuracy characteristics, alignment challenges with established ontologies, and appropriate evaluation methodologies

remains crucial for determining where and how LLM-based extraction can responsibly augment scholarly infrastructure.

Research Questions

This dissertation investigates the application of Large Language Models to automatic extraction of knowledge graphs from Cultural Heritage texts through a systematic examination of existing approaches, the development of novel integrations, and rigorous evaluation across multiple case studies. Three research questions guide this investigation:

> **RQ1: State of Text-to-KG in Cultural Heritage**
>
> What methodologies have Cultural Heritage projects employed to generate knowledge graphs from text, and what domain-specific challenges have these representative projects faced?

> **RQ2: LLMs in text-to-KG pipelines for Cultural Heritage**
>
> How can Large Language Models be integrated into ontology-driven Knowledge Extraction pipelines for Cultural Heritage texts, and what are the limitations, requirements, and trade-offs of such integration?

> **RQ3: LLMs for scholarly opinion mining**
>
> Can LLM-based extraction systems produce Knowledge Graphs of scholarly interpretations that are sufficiently accurate and complete to answer domain expert competency questions while preserving provenance and epistemic uncertainty?

RQ1 addresses the current state of the field through comprehensive survey of existing projects and systematic analysis of documented challenges. Understanding what has been attempted, where bottlenecks persist, and which problems remain unresolved establishes the foundation for methodological development. This question decomposes into three sub-questions: which projects have successfully implemented text-to-KG approaches in Cultural Heritage

contexts, what specific challenges emerged during implementation, and what technological trends in recent literature address these challenges. RQ2 examines how LLMs can augment traditional Knowledge Extraction pipelines given their distinctive capabilities and limitations. Large Language Models offer advantages for tasks requiring semantic understanding with limited training data but introduce new challenges in output standardization, ontological alignment, and prompt engineering. This question investigates the architectural integration patterns, identifies appropriate use cases, and characterizes the trade-offs between traditional supervised approaches and LLM-based methods. The investigation encompasses both theoretical analysis of LLM capabilities and practical methodology development. RQ3 evaluates whether automatically extracted knowledge graphs achieve sufficient accuracy and completeness for scholarly use. This question extends beyond standard Natural Language Processing metrics to examine whether extracted graphs support domain expert competency questions, preserve provenance of extracted information, and represent epistemic uncertainty appropriately. Validation across multiple case studies—scholarly debates about cultural object authenticity, archival finding aids, and academic literature about literary attribution—tests extraction systems under diverse conditions while providing empirical evidence about reliability factors. Through systematic investigation addressing these questions, this work advances practical application of semantic data extraction for cultural institutions while maintaining scholarly standards essential for cultural preservation and interpretation.

Outline

The dissertation is structured to progressively address the three research questions through cumulative development, moving from theoretical foundations and empirical survey to methodological design and applied validation across multiple case studies.

Chapter 1 establishes the theoretical foundations for this work. It dis-

tinguishes between Open and Closed Knowledge Extraction paradigms and situates text-to-KG within Semantic Web frameworks. The chapter provides a formal definition of the text-to-KG task as employed throughout this dissertation and introduces the case studies that serve as empirical testing grounds for the proposed methodologies.

Chapter 2 addresses the first research question by presenting a comprehensive survey of eleven Cultural Heritage projects that implemented text-to-KG methodologies between 2015 and 2025. The chapter categorizes domain-specific challenges identified by project authors, including NLP performance under data scarcity, historical language variation, and entity linking for long-tail entities. The accompanying analysis of 227 publications (2020–2025) examines emerging trends that provide solutions to persistent bottlenecks.

Chapter 3 synthesizes literature-driven design choices supporting LLM application to Knowledge Extraction, examining capabilities and limitations relevant to Cultural Heritage contexts. The chapter establishes theoretical boundaries for LLM deployment, evaluating their strengths in semantic interpretation against limitations in output standardization, annotation reproducibility, and ontological alignment.

Chapter 4 presents a proof of concept demonstrating LLM integration within a text-to-KG pipeline for extracting biographical events from archival finding aids. The chapter validates the alignment of generated knowledge graphs with the RiC ontology through domain expert evaluation.

Chapter 5 introduces *Adaptive Text-to-KG for Cultural Heritage* (ATR4CH), a systematic five-step methodology coordinating LLM-based extraction with ontological frameworks through corpus analysis, Competency Questions, ground-truth annotation development, pipeline architecture design, and multi-metric evaluation.

Chapter 6 presents the primary case study implementing the complete ATR4CH methodology. The chapter describes a pipeline for extracting scholarly debates on cultural object authenticity from Wikipedia articles, employ-

ing the SEBI ontology to represent authenticity assessment interpretations. The chapter presents the full application of the methodology, from preliminary iterations through annotation development to comprehensive evaluation comparing three LLMs across five evaluation dimensions.

Chapter 7 extends the scholarly debate extraction approach to full-length academic articles by applying Retrieval-Augmented Generation to the authorship dispute surrounding *Van den Vos Reynaerde*, a medieval Dutch literary work. The chapter demonstrates how RAG-based architectures address document-length bottlenecks while maintaining alignment with ontological frameworks.

Chapter 8 addresses supporting infrastructure for scholarly opinion mining by presenting a methodology for generating synthetic training data for Aspect-Based Sentiment Analysis. The chapter demonstrates that LLMs can produce large-scale annotation with ontological alignment, and provides benchmarking results from fine-tuning Llama 3.1 8B on domain-specific ABSA tasks.

The Conclusion synthesizes findings to answer the research questions established in this introduction, demonstrating how LLMs can be integrated into ontology-driven Knowledge Extraction pipelines for Cultural Heritage while maintaining scholarly standards. The chapter proposes directions toward configurable infrastructure, methodological refinement, and extended representational capacity for scholarly interpretation and epistemic uncertainty. Supporting materials, including search queries and methodological documentation, are provided in the Appendix.

Chapter 1

Background

This dissertation addresses the challenge of automatically extracting structured knowledge from Cultural Heritage (CH) texts and representing this knowledge as machine-processable graphs conforming to established ontological models. Answering the research questions established in the Introduction requires understanding the theoretical foundations that shape how such systems are designed, evaluated, and deployed within CH contexts.

This chapter establishes these foundations across two interconnected dimensions. First, the Semantic Web (SW) provides the knowledge representation framework that CH institutions have adopted for digital preservation and discovery; understanding how RDF, ontologies, and Linked Open Data technologies enable cross-institutional integration and sophisticated reasoning motivates the focus on ontology-compliant extraction throughout this work. Second, different Knowledge Extraction (KE) paradigms exist for deriving structured information from unstructured text, each with distinct trade-offs; understanding the conceptual differences between Open KE, Closed KE, and fine-grained approaches provides the analytical vocabulary necessary for examining how CH institutions approach content extraction.

Section 1.1 introduces the Semantic Web and Resource Description Framework (RDF), explaining the knowledge representation requirements that shape CH digitization efforts and detailing how SW technologies address distinct requirements across multiple stakeholder communities—researchers, archivists,

scholars, and the general public. Section 1.2 examines KE paradigms from general literature, contrasting Closed KE (ontology-driven extraction within predefined schemas) with Open KE (schema-free extraction discovering structure from text) and fine-grained intermediate approaches that attempt to balance flexibility with semantic rigor. The section also addresses specific requirements for generating SW-compliant output, including URI-based entity identification, data alignment challenges (T-Box, A-Box, and URI alignment), and provenance tracking mechanisms. Section 1.3 synthesizes these foundations and transitions to Chapter 2, which examines how CH institutions have implemented text-to-KG systems in practice through systematic analysis of eleven representative projects.

1.1 Knowledge Representation for Cultural Heritage

The Semantic Web, as articulated by Berners-Lee [22], aims to create an environment where machine-processable information enables software agents to perform sophisticated reasoning, integration, and discovery tasks across distributed data sources. This vision relies on representing knowledge using formal structures that make meaning explicit and computationally tractable. For CH institutions—libraries, archives, museums, and digital humanities initiatives—these representation frameworks address longstanding challenges in cultural preservation, scholarly interpretation, and public access [25].

The Resource Description Framework (RDF)[1] provides the foundational data model for SW applications. RDF represents information as triples of the form $(subject, predicate, object)$, where each element is identified by a Uniform Resource Identifier (URI) that serves as a globally unique identifier. This triple structure enables distributed knowledge representation: different institutions can publish information about the same entities using shared URIs, and automated systems can integrate these descriptions with-

[1] `https://www.w3.org/RDF/`

out requiring centralized coordination. For example, a library might publish the triple (`wd:Q5593`, `dct:creator`, `wd:Q5582`) to express that a particular manuscript was created by Leonardo da Vinci, using Wikidata URIs[2] for both the manuscript (`wd:Q5593`) and the creator (`wd:Q5582`), and Dublin Core[3] vocabulary for the creation relationship (`dct:creator`). Another institution publishing information about Leonardo can reference `wd:Q5582`, enabling automated discovery of related resources across institutional boundaries.

Ontologies provide the semantic layer that gives explicit and machine-readable meaning to RDF data. Specified using the Web Ontology Language (OWL)[4], ontologies define classes (categories of entities), properties (relationships and attributes), and constraints (domain restrictions, cardinality requirements, disjointness assertions) that govern how knowledge can be validly represented within a particular conceptual framework. Domain-specific ontologies developed by CH communities formalize the concepts and relationships necessary for describing cultural resources. The CIDOC Conceptual Reference Model (CIDOC CRM)[5] [51] models museum documentation practices through an event-centric approach, representing cultural objects, actors, places, and the spatiotemporal events that relate them. Dublin Core provides a vocabulary for bibliographic description, defining properties such as `dct:creator`, `dct:date`, and `dct:subject` widely adopted across library systems. PROV-O[6] models provenance, documenting how entities derive from other entities and which agents and activities participated in their creation. Additionally, controlled vocabularies such as the Getty Vocabularies[7] provide standardized terminologies for concepts like art historical periods, geographic locations, and cultural object types, enabling precise semantic annotation across diverse domains [88].

CH institutions have invested decades of curatorial expertise in developing these ontological frameworks, motivated by several imperatives. First, shared

[2] `https://www.wikidata.org/`
[3] `https://dublincore.org/`
[4] `https://www.w3.org/OWL/`
[5] `http://www.cidoc-crm.org/`
[6] `https://www.w3.org/TR/prov-o/`
[7] `http://www.getty.edu/research/tools/vocabularies/`

ontologies enable cross-institutional discovery and integration that would be impossible with institution-specific metadata schemas. Second, formal ontologies support sophisticated reasoning through their grounding in description logic [167]; an automated system, based on CIDOC CRM, can infer that if an object is a `crm:E22_Man-Made_Object` and participates in a `crm:E12_Production` event, then the object has a temporal beginning corresponding to its production date, even if that date is not explicitly stated in the original metadata. Third, ontology-based representations preserve scholarly interpretation by making conceptual commitments explicit [199]; when a cataloger asserts that an artifact instantiates the class `crm:E84_Information_Carrier` rather than merely `crm:E22_Man-Made_Object`, this assertion captures domain expertise about the artifact's primary function and cultural significance.

The combination of RDF and domain ontologies has achieved notable successes in CH metadata publication. Europeana[8], which aggregates digital collections from thousands of European institutions, employs SW technologies to enable cross-collection discovery and integration. The Digital Humanities community has engaged extensively with these infrastructures, developing tools, methodologies, and applications for cultural analysis, preservation, and public participation [193, 201].

The adoption of SW technologies addresses distinct requirements across multiple stakeholder communities in CH contexts. For researchers, ontology-based representations enable pattern discovery across collections through queries that identify relationships invisible in traditional catalog systems: tracing biographical networks by connecting individuals through shared activities, mapping the geographic mobility of cultural objects across institutional holdings, or analyzing the evolution of artistic styles by examining temporal and influence relationships between works [97]. For archivists and CH professionals, RDF-based systems provide advantages over traditional database architectures through flexible schema evolution that accommodates emerging descriptive re-

[8] `https://www.europeana.eu/`

quirements without costly database migrations, distributed authority control that enables collaborative description across institutions while maintaining referential integrity, and native support for representing complex provenance chains that document custodial histories, authenticity assessments, and the evidential basis for attribution claims [38]. For scholars, these technologies support computational discourse analysis by enabling queries that examine how interpretative frameworks, terminologies, or scholarly arguments propagate through citation networks and evolve across temporal contexts, preserving the epistemic stance and evidential reasoning that characterize humanistic inquiry [40]. For the general public, Linked Open Data enables intuitive discovery interfaces that connect resources through meaningful relationships—exploring artworks depicting related subjects, manuscripts produced in the same scriptorium, or biographical connections between historical figures—without requiring knowledge of technical metadata schemas or institutional cataloging practices [115].

However, this large-scale process has focused predominantly on publishing information *about* resources—the metadata that describes what resources are, when they were created, who was involved in their creation, and where they are held. For textual resources, the *content* of the text itself—the entities mentioned, relationships expressed, events described, and arguments constructed—has received less systematic attention in SW publication efforts [78]. While cataloging metadata might record basic bibliographic information about a letter, extracting the content knowledge that this letter discusses specific concepts or disputes requires different methodologies capable of processing the natural language text contained within the resource itself.

Extending KG extraction to textual content yields benefits across multiple dimensions. For researchers, content-based querying enables discovery patterns impossible with metadata alone: identifying all documents discussing particular historical events, tracing how specific arguments evolve across a scholarly corpus, or discovering implicit connections between concepts mentioned in dif-

ferent sources. For archivists, automated content extraction supplements manual cataloging for large collections where detailed subject indexing is resource-prohibitive, enabling discovery through document content rather than externally assigned subject headings. For digital humanities scholars, structured representations of textual content enable computational analysis of discourse patterns, argument structures, and knowledge circulation across large corpora, supporting research questions about intellectual history, conceptual change, and the social dynamics of scholarly communication. For the general public, content-based access lowers barriers to CH engagement by supporting natural language queries that retrieve documents based on what they discuss rather than requiring familiarity with cataloging vocabularies or archival organization principles.

The challenge, then, lies in developing methodologies that can automatically extract structured knowledge from natural language text and represent this knowledge in forms compatible with the ontological frameworks and SW standards that CH institutions have adopted for metadata. The following sections examine the KE paradigms developed to address this challenge, focusing on approaches that align with institutional practices and establish evaluation criteria for text-to-KG systems.

1.2 Knowledge Extraction Paradigms

Knowledge Extraction encompasses computational approaches for deriving structured information from unstructured sources. Within this field, two paradigms have emerged based on whether extraction operates within predefined schemas or discovers structure directly from text. Understanding these paradigms and their respective trade-offs provides the conceptual foundation for analyzing how CH institutions approach content extraction from textual resources.

1.2.1 Closed Knowledge Extraction

Closed Knowledge Extraction (Closed KE) operates within predefined ontological frameworks that specify permitted entity types, relations, and semantic constraints [95]. Systems following this paradigm extract information conforming to target schemas expressed in formal knowledge representation languages such as OWL and formats such as RDF. When the extraction process is explicitly guided by ontologies that define classes, properties, and constraints, this approach is often termed Ontology-Based Information Extraction (OBIE) [215].

The Closed KE paradigm aligns naturally with the SW vision articulated by Berners-Lee [22], as extraction targets conform to ontological schemas designed for machine reasoning and cross-system integration. Within contexts where domain-specific ontologies provide the semantic foundation for institutional knowledge organization, Closed KE enables automated content extraction to produce outputs directly compatible with established metadata infrastructures. For example, a library implementing Closed KE to process manuscript descriptions can extract entities and relationships that instantiate the same classes and properties used in their existing catalog records, enabling seamless integration of automatically extracted content with manually curated metadata.

Closed KE systems typically decompose the extraction task into several specialized subtasks, each addressing a distinct aspect of the overall challenge. Named Entity Recognition (NER) identifies entity mentions in text and classifies them according to predefined types such as Person, Location, Organization, and domain-specific categories like ArtWork, ChemicalSubstance, or HistoricalEvent. Entity Linking, also termed Entity Resolution or Entity Disambiguation, maps entity mentions to canonical identifiers in authority files or knowledge bases, resolving ambiguity when the same surface form can refer to multiple distinct entities; for example, distinguishing "Paris" the city from "Paris" the figure in Greek mythology, or linking "Leonardo" mentions to the

Wikidata URI for Leonardo da Vinci rather than to other historical figures sharing that name. Relation Extraction identifies semantic relationships between entities that correspond to ontology properties, such as extracting that Newton *influenced* Leibniz or that a particular artwork was *created_in* Florence during 1504. Event Detection, particularly important when using event-centric ontologies like CIDOC CRM, identifies occurrences of events described in text and extracts their participants, locations, and temporal boundaries. Finally, these extracted components must be assembled into coherent graph structures, with validation ensuring that all assertions conform to ontological constraints such as property domain and range restrictions.

The primary advantages of Closed KE stem from its alignment with existing knowledge organization systems and institutional practices. First, ontological constraints provide quality assurance by rejecting extractions that violate domain semantics; a system enforcing CIDOC CRM constraints cannot assert that a `crm:E39_Actor` has a `crm:P108_has_produced` relationship to another `crm:E39_Actor`, as this property requires the range to be a `crm:E24_Physical_Man-Made_Thing`. Second, extracted KGs inherit the formal semantics of their governing ontologies, supporting post-processing through reasoning; if the ontology specifies that `crm:P108_has_produced` is transitive, a reasoning engine can infer derived relationships not explicitly extracted from text. Third, the predefined schema facilitates evaluation by providing clear criteria for correctness; an extraction asserting that Leonardo created the Mona Lisa can be validated against the target ontology's definition of the `dct:creator` property and ground truth annotations specifying the correct relationship for that entity pair.

However, Closed KE traditionally requires substantial domain expertise and labeled training data. Developing or selecting appropriate ontologies demands understanding of the domain's conceptual structure and scholarly conventions. Adapting general-purpose NER systems to recognize domain-specific entity types typically requires annotated training examples that identify entity

boundaries and types in sample texts. Training Relation Extraction models requires corpora annotated with relationship instances conforming to the target ontology's property definitions. These resource requirements have historically limited Closed KE deployment to domains where institutions could invest in ontology engineering, corpus annotation, and system customization.

1.2.2 Open Knowledge Extraction

Open Knowledge Extraction (Open KE), also termed Open Information Extraction (Open IE), operates without predefined schemas or ontological constraints [221, 64]. Systems following this paradigm extract facts from text in the form of relation tuples (e_1, r, e_2), where entities e_1 and e_2 are connected by relation r, without requiring prior specification of entity types or relation taxonomies. The extraction process is driven by linguistic patterns and syntactic structures rather than domain-specific knowledge.

The TextRunner system [221], among the first implementations of this paradigm in 2007, performed web-scale extraction of relation tuples directly from surface text forms. For example, from the sentence "Thomas Edison invented the light bulb in 1879", TextRunner would extract the tuple $(\text{Thomas Edison}, \text{invented}, \text{light bulb})$ by recognizing the subject-verb-object pattern, without consulting any knowledge base to determine that "Thomas Edison" refers to a specific historical figure or that "invented" represents a particular type of relationship defined in some ontology. Other notable Open KE systems include NELL (Never-Ending Language Learner) [31], which performs continuous, unsupervised extraction from web text while incrementally refining its extraction patterns through self-supervised learning, and OpenIE [131], which uses dependency parsing to identify extractable relationships with minimal linguistic assumptions.

The primary advantage of Open KE lies in its domain independence and scalability. A single Open KE system can process documents across diverse topics without requiring domain-specific training data, ontology engineering, or manual adaptation. This generality has proven valuable for web-scale knowl-

edge base construction projects such as Google's Knowledge Vault and Microsoft's Satori, where the goal is extracting broad coverage of factual information rather than populating a specific domain ontology.

However, Open KE faces several limitations when applied to contexts requiring integration with formal knowledge representation frameworks. First, entity representation lacks standardization; multiple extractions may refer to the same entity using different surface forms ("Leonardo da Vinci", "Leonardo", "da Vinci") without recognizing these as coreferential, leading to entity proliferation in the extracted knowledge base. Second, relation extraction produces surface-level predicates tied to specific linguistic expressions rather than normalized semantic relationships; "created", "painted", "produced", and "made" might represent semantically equivalent relationships in a target domain, but Open KE systems treat them as distinct predicates. Third, extracted facts lack the semantic grounding provided by ontological definitions, complicating tasks such as consistency checking, subsumption reasoning, or cross-domain integration. Fourth, quality control mechanisms are limited; without domain-specific constraints, Open KE systems cannot distinguish plausible but incorrect extractions from valid assertions based on semantic compatibility.

For institutions maintaining knowledge organization systems based on formal ontologies and controlled vocabularies, these limitations create barriers to adopting Open KE outputs. The gap between surface-level extraction tuples and the semantically grounded, URI-identified assertions required for Linked Open Data publication necessitates extensive post-processing, entity resolution, and manual validation.

1.2.3 Fine-Grained Approaches

Recent systems attempt to combine Open KE's flexibility with Closed KE's semantic rigor by employing intermediate representations that maintain ontological grounding while requiring fewer domain-specific assumptions during text processing. These approaches aim to reduce the manual engineering asso-

ciated with traditional Closed KE while preserving compatibility with formal knowledge representation frameworks.[9]

Text2AMR2FRED represents one such approach, transforming natural language text into Abstract Meaning Representation (AMR) graphs and subsequently mapping these linguistic structures to formal ontologies through the FRED (Formal Representation of Events and Descriptions) framework.[10] AMR [14] provides a graph-based semantic representation that abstracts away from surface syntax while capturing predicate-argument structure, named entities, and semantic roles. The subsequent mapping to ontology vocabularies enables integration with SW infrastructures while avoiding the need for domain-specific extraction rules targeting individual relation types. However, this approach introduces alignment challenges; the linguistic categories implicit in AMR structures do not necessarily correspond cleanly to the conceptual distinctions encoded in domain ontologies like CIDOC CRM, requiring complex mapping rules and potentially losing semantic nuances in the translation process.

ReLiK (Retrieve, Read and LinK)[11] implements a different architectural strategy, decomposing KE into coordinated retrieval and reading components [145]. The retriever employs dense encoders to identify relevant candidate entities or relations from large knowledge bases, effectively narrowing the search space for the downstream reader. The reader then performs span extraction for Entity Linking or triplet extraction for Relation Extraction, selecting from retrieved candidates. This two-stage design enables processing over extensive entity vocabularies, such as all Wikipedia entities, while maintaining computational efficiency. ReLiK achieves performance approaching or exceeding traditional supervised systems on established benchmarks—86.4 F_1 on AIDA for Entity Linking and 95.0 F_1 on NYT for Relation Extraction—while sup-

[9] This dissertation focuses on Closed Knowledge Extraction approaches. However, a given implementation within a Closed KE framework may incorporate elements from Open or Fine-Grained KE, as the boundaries between these categories are not always rigid in practice.

[10] `http://wit.istc.cnr.it/stlab-tools/fred/`

[11] `https://github.com/SapienzaNLP/relik`

porting both entity resolution to external knowledge bases and extraction with respect to predefined relation schemas. Having access to entity resolution, extracted relations can be mapped to ontological properties to produce RDF graphs, and the model supports fine-tuning for domain-specific relation types.

1.2.4 Knowledge Extraction for the Semantic Web

Adapting Closed KE to produce SW-compliant output requires addressing requirements beyond general ontology conformance [95]. While traditional Information Extraction might identify that "Leonardo" is a Person and that this Person created "Mona Lisa", generating RDF suitable for LOD publication imposes additional constraints on representation and linking.

First, entity mentions must be mapped not merely to type labels but to URIs that serve as globally unique identifiers within the LOD cloud. This requirement elevates Entity Linking from an optional refinement to a fundamental component of the extraction pipeline. An institution cannot publish the extracted assertion that Leonardo created the Mona Lisa as LOD unless both entities are identified by URIs—ideally linking to authority files such as Wikidata (`wd:Q762` for Leonardo, `wd:Q12418` for the Mona Lisa) or institutional collection databases. When entities mentioned in texts lack entries in large-scale knowledge bases, institutions must mint local URIs following LOD best practices, maintaining alignment between local identifiers and institutional metadata systems.

Second, [95] identify three dimensions along which data alignment challenges manifest in SW-oriented extraction systems. T-Box alignment concerns the ontology schema itself; when source texts mention entity types or relationships not captured in the target ontology, the system must either extend the schema with new classes and properties or acknowledge extraction limitations. A-Box alignment addresses instance-level challenges; when multiple extractions refer to the same entity using different surface forms or when the same entity appears in both automatically extracted data and manually curated catalog records, deduplication and consistency checking become essential. URI

alignment focuses on entity resolution; determining when entity mentions correspond to existing knowledge base entries versus representing novel entities requires sophisticated disambiguation considering context, co-referent mentions, and external evidence.

Third, provenance tracking and validation mechanisms must ensure that generated RDF conforms to both syntactic requirements (well-formed RDF syntax) and semantic constraints (adherence to property domains, ranges, and cardinality restrictions defined in the ontology). Systems may employ constraint languages such as SHACL[12] to validate generated graphs or integrate validation into the extraction process itself, rejecting candidates that would violate ontological constraints.

1.3 Chapter Summary

This chapter established the foundational knowledge representation frameworks and extraction paradigms necessary for understanding text-to-KG approaches in CH contexts. Section 1.1 introduced the Semantic Web and RDF as the knowledge representation frameworks adopted by CH institutions for digital preservation and discovery, explaining how ontology-based systems enable cross-institutional integration, sophisticated reasoning, and preservation of scholarly interpretation. The observation that SW publication efforts have focused predominantly on metadata rather than textual content motivates the extraction methodologies examined throughout this dissertation. Section 1.1 also detailed how SW technologies address distinct requirements across multiple stakeholder communities—researchers, archivists, scholars, and the general public—each benefiting from ontology-driven knowledge organization in different ways.

Section 1.2 contrasted Open and Closed KE paradigms, clarifying the conceptual and practical trade-offs between schema-free extraction and ontology-driven approaches. While Open KE offers domain independence and scalabil-

[12] `https://www.w3.org/TR/shacl/`

ity, Closed KE provides the ontological grounding, entity standardization, and semantic constraints necessary for integration with institutional knowledge organization systems. Fine-grained approaches such as Text2AMR2FRED and ReLiK represent attempts to balance flexibility with semantic rigor. The subsection on KE for the Semantic Web identified additional requirements that SW-compliant extraction must satisfy: URI-based entity identification, T-Box/A-Box/URI alignment, and provenance tracking with validation mechanisms.

Having established these theoretical foundations from general literature on KE and SW technologies, the following chapter examines how CH institutions have implemented text-to-KG systems in practice. Chapter 2 presents a systematic survey of eleven representative projects that have generated KGs from CH texts, analyzing their methodological choices, identifying domain-specific challenges that distinguish CH text processing from general-domain extraction, and synthesizing architectural patterns and trends from recent literature. This empirical analysis reveals which extraction paradigms CH institutions adopt, which technical components present persistent bottlenecks, and which methodological innovations address the resource constraints and quality requirements characteristic of CH applications.

Chapter 2

Text-to-KG in the Cultural Heritage domain

Chapter 1 established the theoretical foundations necessary for understanding text-to-KG extraction, including the knowledge representation frameworks adopted by CH institutions (RDF, ontologies, Linked Open Data) and the distinction between Open and Closed Knowledge Extraction (KE) paradigms. This chapter[1] addresses **RQ1**: *What methodologies have Cultural Heritage projects employed to generate knowledge graphs from text, and what domain-specific challenges have these representative projects faced?*

Through systematic analysis of eleven representative CH text-to-KG implementations developed between 2015 and 2025, this chapter examines how theoretical paradigms manifest in practical systems, identifies domain-specific challenges that distinguish CH text processing from general-domain extraction, and synthesizes methodological patterns observed across projects and recent literature. The analysis proceeds from empirical observation to theoretical synthesis: Sections 2.2 through 2.5 document what CH projects have implemented and which obstacles they encountered, while Section 2.6 distills these observations into a formal framework characterizing text-to-KG approaches in CH contexts.

[1]This chapter draws from prior work by the author: Schimmenti, A., Vitali, F., & van Erp, M., *Knowledge Graph Extraction in the Cultural Heritage domain: A Survey of Projects, Challenges and Trends* (2025, to be submitted to *Journal of Web Semantics*).

2.1 Background

Although digitization efforts in recent decades have successfully published metadata through SW technologies, the content of CH documents, including archives, scholarly texts, manuscripts, and historical narratives, remains largely inaccessible through semantic querying [78, 38]. Institutional metadata schemas provide structured access through standardized fields, yet relationships between concepts, people, places, and events frequently remain implicit within unstructured textual descriptions [78]. Manual creation of structured representations from these texts remains resource-intensive, particularly given the volume of unstructured text in CH collections, driving increased research into automatic and semi-automatic KG construction techniques. Recent methodological attention to text-to-KG extraction is evidenced by dedicated workshop series (the International Workshop on Knowledge Graph Construction reached its fifth edition in 2025 [196]), specialized challenges (Knowledge-Base Construction from Pretrained Language Models Challenge at ISWC 2024 [126]), and journal special issues [107, 213]. However, mainstream KE datasets typically emphasize Wikipedia, product reviews, and news articles rather than the heterogeneous documents that constitute CH corpora—archival materials, manuscripts, letters, and curated artifact descriptions [197]. Existing literature provides complementary perspectives on related areas. Maynard et al. [132] survey NLP approaches for the SW, and Rodriguez et al. [95] examine specifically information extraction techniques techniques. Ranjgar et al. [161] focus on data modeling and publishing in CH contexts, and Tamašauskaitė and Groth [188] survey KG development processes broadly. Regino et al. [163] analyze challenges in KG extraction for SW applications. This chapter complements these works by focusing specifically on text-to-KG extraction implementations within CH domains. To address **RQ1** systematically, this chapter investigates three interrelated sub-questions:

- **RQ1.1.** Which projects have successfully generated KGs from CH texts using text-to-KG methodologies?

- **RQ1.2.** What challenges emerge when applying KE techniques to CH documents?

- **RQ1.3.** What methodological trends address the challenges identified in CH text-to-KG implementations?

The analysis employs a three-stage approach: an empirical case study analysis to identify implemented solutions (RQ1.1), extraction of domain-specific challenges from these implementations (RQ1.2), and a trend analysis across broader literature to identify methodological patterns (RQ1.3). Section 2.2 details the survey methodology. Section 2.3 analyzes eleven projects implementing text-to-KG approaches. Section 2.4 examines CH-specific obstacles requiring specialized approaches. Section 2.5 identifies methodological patterns across surveyed projects and related literature. Section 2.7 synthesizes findings and implications.

2.2 Methodology

This systematic review employed a two-phase approach: (1) identification of CH text-to-KG projects through diverse search strategies, and (2) literature analysis to identify methodological trends addressing identified challenges.

2.2.1 Project Identification Strategy

The research employed five complementary search methods to ensure comprehensive coverage:

- **Scholarly Database Searches:** Structured queries across Google Scholar [80], Elsevier ScienceDirect [62], Springer [184], ACM Digital Library [1], IEEE Xplore [100], and DBLP [42] combined domain-specific and technical terms: "knowledge graph extraction," "cultural heritage," "named entity recognition," "relationship extraction," "GLAM," "ontology population," and "digital humanities." Representative Google Scholar queries yielded: "Knowledge Graph" AND "Digital Edition" (140 items);

"Knowledge Extraction" AND "Digital Edition" (42 items); "Information Extraction" AND "Knowledge Graph" AND "Humanities" (2,120 items).

- **Venue-Specific Analysis:** Manual review of abstracts from conferences and journals at the intersection of Semantic Web and Digital Humanities research, including ISWC [103], ESWC [63], ADHO [2], AIUCD [4], DHBenelux [46], EKAW [58], Semantic Web Journal [101], Journal of Web Semantics [61], and specialized workshops including Text2KG [195], SemDH [182], and LM-KBC [125].

- **Directory Exploration:** Systematic searches of Digital Humanities directories including the Catalogue of Digital Editions [49] and Digital Humanities Atlas [45].

- **Citation Network Analysis:** Forward citation analysis using Open Citations [144] on seminal papers identified through initial searches.

- **Community Engagement:** Consultation with Digital Humanities communities through professional networks and mailing lists, particularly the AIUCD forum [3].

The analysis focused on projects from 2015-2025 to capture multi-year development cycles typical of text-to-KG implementations. From the initial pool, four inclusion criteria were applied:

1. **Methodological transparency**: Detailed documentation of text-to-KG approaches

2. **CH domain focus**: Implementation targeting the automatic extraction of KGs

3. **Demonstrable implementation**: Actual deployment evidenced by e.g. downloadable datasets, accessible interfaces, or detailed implementation reports

4. **Evaluation documentation**: Standard evaluation metrics (e.g. precision, recall, F_1-score) or domain-specific measures documented in publications or project reports

Projects with well-documented approaches but limited reports on the results (e.g. ongoing projects) were included if the rest of the criteria were satisfied. A structured analytical framework capturing eight dimensions enabled systematic comparison (Table 2.1). Eleven projects met all inclusion criteria and received detailed analysis.

Table 2.1: Dimensions of Analysis for Text-to-KG Approaches

Dimension	Guiding Question
Project	What are the project's scope and objectives?
Input	What types of documents are processed?
Methodology	Which extraction techniques are employed?
Representation	Which ontologies and vocabularies are used?
Evaluation	How is extraction quality assessed?
Output	What KGs are produced (size, format)?
Challenges	What domain-specific obstacles were encountered?
Distinctive Features	What novel approaches were introduced?

Projects are organized by primary data type: (1) **Structured and Semi-structured Sources** with consistent organizational patterns enabling rule-based extraction (e.g., HyperReal processing Steven Olderr's *Symbolism: a Comprehensive Dictionary* [174]); (2) **Unstructured Narrative Texts** requiring sophisticated NLP pipelines for implicit relationships (e.g., Viewsari processing Vasari's *The Lives* [142]); (3) **Specialized Document Types** with domain-specific terminologies or historical language variants (e.g., Notarypedia processing medieval legal documents [60]).

2.2.2 Literature Analysis for Trend Identification

To identify methodological trends, literature analysis targeted publications from 2020-2025, reflecting rapid evolution following advances in transformer-based language models. Three complementary strategies were employed: First, traditional searches using the same databases as 2.2.1, across five thematic

categories: (1) primary methodological queries ("knowledge graph extraction" AND "cultural heritage"), (2) data alignment challenges ("entity disambiguation" AND "historical texts"), (3) domain-specific extraction ("historical language" AND "NLP"), (4) document phenomenology ("literary language" AND "knowledge extraction"), and (5) methodological approaches such as ("BERT" AND "knowledge extraction" AND "humanities"). The complete query list appears in Appendix.

Second, 15 highly cited papers (30+ citations) including foundational surveys [132] were selected, and the Open Citations Index was used to identify citing papers, based on the assumption that current text-to-KG research would reference foundational works. Additional scripts were used to clean the resulting lists and perform analysis on the titles, which are available as a GitHub repository.[2]

Third, bibliographic references from the projects analyzed were incorporated as an additional source. These approaches yielded approximately 1,000 papers. The selection of title and abstract identified approximately 200 articles discussing text-to-KG or relevant subtasks for detailed analysis. These results are discussed in Section 2.5.

2.3 Case Studies

This section analyzes eleven projects that implement text-to-KG extraction between 2015-2025 and that meet established inclusion criteria. Table 2.2 provides an overview organized by primary data type. Detailed comparison tables for each project , appear in Appendix.

The analyzed projects span various CH subdomains, including music history, visual arts, literary studies, historical archives, and museum collections. Each project is discussed in turn, categorized by primary methodological approach and data type.

[2] `https://github.com/aschimmenti/citation_trend_analysis_survey.git`

Table 2.2: Overview of Text-to-KG Projects in Cultural Heritage

Project	Years	Domain
Structured and Semi-structured Sources		
BiographySampo	2018-present	Finnish biographical heritage
ParliamentSampo	2020-2023	Finnish parliamentary records
WarSampo	2014-present	WWII Finnish military history
HyperReal	2021	Cultural symbolism
Unstructured Narrative Sources		
Odeuropa	2020-2023	European olfactory heritage
MMKG	2022-2025	Musical encounters
MusicBO	2022-2025	Bologna musical heritage
Viewsari	2024	Renaissance art history
Bernoulli-Euler	2016-2023	Scientific correspondence
Specialized Document Types		
InTaVia	2021-2024	European biographical integration
Notarypedia	2018-2019	Medieval legal documents

2.3.1 Structured and Semi-Structured Sources

Sampo

Sampo constitutes both a model and a series of semantic portals for publishing and integrating multiple LOD datasets, services, and portals of Finnish CH [98]. Across different Sampo implementations, several key tools and approaches are shared:

- Secompling: Used for KE and linguistic analysis
- FinBERT: A Finnish language version of BERT for deep learning-based text processing
- SPARQL ARPA: A configurable automatic annotation tool using lexical analysis, SPARQL, and ontologies to identify entities
- AATOS: An annotation framework linking entities to ontologies and ranking candidate terms
- LAS (Lexical Analysis Services): Provides linguistic processing capabilities

The Sampo approach consistently emphasizes publishing extracted knowledge as linked data through SPARQL endpoints, enabling flexible querying

and integration across datasets and applications. This infrastructure supports faceted search interfaces, visualization tools, and analytical capabilities transforming raw textual sources into semantic knowledge networks [98]. Three projects within the Sampo series leverage NLP and semantic technologies to transform diverse textual sources into KGs.

BiographySampo BiographySampo[3] transforms biographical texts into Linked Open Data for prosopographical research [99]. The system processes over 13,600 life stories from the Biographical Centre of the Finnish Literature Society, converting them into a KG published through a SPARQL endpoint and semantic portal. Table 8.6 summarizes key characteristics of BiographySampo.

The project employs Bio CRM,[4] an extension of CIDOC CRM, as its foundational data model with `:Person` as the central class. The transformation pipeline features two complementary branches: pattern-based extraction using regular expressions for semi-structured portions (family relations, career events) and an NLP pipeline for free text processing. The NLP component applies Finnish dependency parsing with CoNLL-U[5] output, transforms linguistic data into RDF using CoNLL-RDF, and employs symbolic reasoning to establish connections between sentences, words, and document structure.

Manual evaluation of 135 events yields 99% precision for event identification, 98% accuracy for temporal information, and 98% precision with 77% recall for place linking. The system combines pattern-based extraction with statistical reasoning about historical lifespans and family relationships, demonstrating a particular strength in processing the formulaic structure of biographical summaries.

The resulting KG contains approximately 10 million triples representing 13,665 biographies, enriched through data linking to 16 external sources including Wikidata, VIAF, and specialized Finnish CH databases.

[3] `https://www.ldf.fi/dataset/nbf`
[4] `http://ldf.fi/schema/bioc/`
[5] `https://universaldependencies.org/format.html`

ParliamentSampo ParliamentSampo[6] processes approximately one million speeches from Parliament of Finland plenary sessions from 1907-2024 [189]. The project, funded by the Academy of Finland's DIGIHUM program from 2020-2022, converts parliamentary debates into RDF triples for computational analysis. Table 8.7 summarizes key characteristics of ParliamentSampo.

The project uses OCR to digitize parliamentary minutes. The pipeline applies linguistic annotation and automatic content analysis to speeches from 2015 onwards. The extraction process identifies speakers, political affiliations, speech types, dates, and procedural elements. Each speech links to its original parliamentary document, and recent speeches include links to video recordings.

The ontology construction process extracts classes and properties from Parliament databases rather than using predefined vocabularies, mapping the data origin structure. The system models members of parliament, political groups, committees, legislation, and parliamentary procedures. NIF represents linguistic data, Dublin Core Terms provides metadata, and CIDOC CRM models entity relationships.

The portal[7] provides seven application views: faceted search, temporal analysis, network visualization, statistical analysis, and geographic mapping. Users can filter speeches by content, speaker, party, date, and document type. The system tracks speaker networks and political relationships over time. No explicit evaluation process is documented.

WarSampo WarSampo[8] processes World War II historical data from Finland covering the Winter War (1939-1940), Continuation War (1941-1944), and Lapland War (1944-1945) [190]. The system launched in 2015 and contains 14 million triples representing 95,000 war casualties, 26,400 war diaries, 164,000 photographs, and 3,360 magazine articles. Table 8.8 summarizes key characteristics of WarSampo.

The project integrates data from 20 sources including the National

[6] `https://seco.cs.aalto.fi/projects/semparl/en/`
[7] `https://parlamenttisampo.fi/`
[8] `https://www.sotasampo.fi/en/`

Archives of Finland, Finnish Defence Forces, and National Land Survey of Finland. AATOS, a configurable annotation tool, processes magazine articles from *Kansa Taisteli*, which contains memoirs from Finnish military personnel and civilians. The annotation pipeline includes text extraction from PDFs using OCR tools (ABBYY FineReader and Tesseract), linguistic preprocessing to transform words into base forms, entity linking using the SPARQL ARPA tool with multiple domain ontologies, and candidate ranking based on configurable parameters.

WarSampo employs CIDOC CRM as its data model foundation with domain-specific ontologies covering people, military units, Karelian places, municipalities, and external resources such as DBpedia.[9] The system creates homepages for deceased soldiers and 5,000 surviving soldiers, military units, places, and events. Each homepage aggregates information from multiple sources including war diaries, photographs, and memoir articles.

Evaluation on 433 articles demonstrated precision up to 82.02% and recall up to 75.76% for military unit identification. OCR post-processing significantly improved results across all entity types.

HyperReal

HyperReal constitutes a KG dedicated to cultural symbolism. Sartini et al. [174] address a gap in the Semantic Web context by providing structured representations of symbolic relationships between cultures. Table 8.9 summarizes key characteristics of HyperReal.

The project is built upon the Simulation Ontology [174], specifically developed to model background knowledge of symbolic meanings. The KG was created through reengineering symbolic knowledge from heterogeneous resources and mapping it into the Simulation Ontology schema. The ontology incorporates reusable ontology patterns including situation, semiotics, and information realization to provide a foundation for representing symbolic relationships.

Data were extracted and converted from three main sources: DBpedia,

[9] `https://www.dbpedia.org/`

WordNet,[10] and Olderr's dictionary of symbols [139], resulting in more than 41,000 simulations captured in nearly 500,000 triples [174].

HyperReal expresses context-dependent symbolic meanings. For example, it represents how an owl symbolizes death in Hindu, Japanese, and Mayan contexts, while representing helpful spirits in Siberian contexts. The ontology includes specialized properties to express different types of symbolic relationships, such as protection simulations (where a simulacrum symbolically protects against something) or healing simulations (where a simulacrum symbolically cures something). HyperReal showcases a standard case for semistructured data KE, where even if the documents are in plain text, the syntax is formulaic enough to fit a simple pattern matching structure. On average, the KE achieves a mean 0.97 F_1 across all nodes of the KG, with errors being caused by variations in the dictionary entries.

2.3.2 Unstructured Narrative Sources

Odeuropa

Odeuropa is a H2020 European initiative[11] developed an approach to extract olfactory experiences from historical texts, representing sensory heritage using KGs [121]. The project analyzed 167,029 textual resources in six European languages (English, Italian, French, Dutch, German, and Slovenian) and 43,679 visual materials, resulting in 2.4 million smell instances in the European Olfactory KG. The Odeuropa Smell Explorer[12] provides an interface to navigate the KG through three dedicated paths: smell sources (550 cataloged sources), fragment spaces (115 documented spaces), and gestures and allegories (35 olfactory gestures). Table 8.11 summarizes key characteristics of Odeuropa.

The project applied a "KG annotation" approach, where the annotation model to represent smell-related events [203] was developed to be exactly mapped to an ontology [122]. This methodology comprises two parallel de-

[10] `https://wordnet.princeton.edu/`
[11] `https://odeuropa.eu`
[12] `https://explorer.odeuropa.eu/`

velopment tracks:

- **Corpus Annotation Model:** The project developed a specialized frame-based annotation scheme inspired by FrameNet but specifically tailored to capture olfactory situations in text. This annotation framework defined nine Frame Elements (FEs): `smell source`, `quality`, `perception`, `evoked smell`, `location`, `time`, `effect`, `odor carrier`, and `circumstances` to capture smell-related events. The training corpus contained approximately 1,700 manually annotated olfactory events per language, with an imbalanced distribution favoring Smell Source and Quality as core elements.

- **Ontology Development:** In parallel, Odeuropa developed a three-layer ontological model based on CIDOC CRM and CRMsci,[13] with specific extensions for sensory and olfactory information. The ontology followed a user-centered design approach guided by over 70 competency questions from domain experts.

The technical pipeline used multilingual BERT-based models trained using a multitask learning approach, where each element of the frame was treated as a separate task. The system utilized language-specific pre-trained models (bert-base-cased for English, bert-base-italian-cased for Italian, flaubert-base-cased for French, etc.) alongside multilingual BERT for cross-lingual analysis. The multitask framework consistently outperformed single-task classification, demonstrating the effectiveness of treating frame element detection as interconnected but distinct tasks.

The result is a seamless pipeline that starts from annotation, to training and extraction, until the mapping to LOD. For example, when the text extraction system identified a Smell Word and its associated Smell Source and Quality in a historical document, these elements could be directly assigned to the corresponding ontological concepts: `L11 Smell`, `L12 Smell Emission`

[13] `https://cidoc-crm.org/crmsci`

with `F3 source`, and `Attribute Assignment` with appropriate assigned values.

The evaluation methodology demonstrated varying performance between frame elements, with core elements achieving higher accuracy (Smell Words: 65.5-88.7% F_1, Smell Source: 23.5-57.5% F_1, Quality: 44.3-80.1% F_1) compared to peripheral elements such as Effect and Time, which suffered from boundary detection issues and limited training instances. Cross-lingual analysis revealed that multilingual training was beneficial only for certain languages (Italian, German, Dutch), while others performed better with monolingual models.

The project addressed significant challenges in the processing of subjective sensory information, including cross-cultural variations in smell terminology, historical language changes, and the disambiguation of polysemous smell-related words. The system tackled the inherently subjective nature of olfactory descriptions by developing standardized taxonomies that could accommodate cultural and temporal variations in smell perception and description.

The Smell Explorer interface enables "nose-first" querying, allowing users to search by smell sources rather than traditional metadata categories. The platform includes specialized features such as interactive nosebooks for olfactory analysis, temporal visualization of smell descriptions, geographic mapping of smell references, and word clouds summarizing olfactory characteristics. Users can export search results and access the underlying data through SPARQL endpoints for advanced computational analysis.

Polifonia

The Polifonia project[14] is a H2020 European initiative focused on preserving and connecting diverse elements of musical heritage. The project aims to address the challenge of describing, connecting, and preserving heterogeneous traces of musical expressions and experiences across Europe. Two of its pilots, MMKG and MusicBO, demonstrate complementary KE approaches from mu-

[14] `https://polifonia-project.eu/`

sical heritage texts, sharing a common network of ontologies for representation. [21]

MMKG Musical Meetups Knowledge Graph (MMKG) [137] used a text-to-KG pipeline on over 33,309 Wikipedia biographies of musical artists. The project focuses on meetups, contacts and interactions in European musical culture between 1800 and 1945, resulting in 45,812 historical meetups involving 49,170 people, 7,107 places, and 51,120 temporal expressions within the final KG. Table 8.13 summarizes key characteristics of MMKG.

MMKG is built on the MEETUPS ontology, which models historical encounters as structured named graphs describing: participants, locations, temporal expressions, and purposes. The ontology incorporates established vocabularies including Time Ontology, PROV Ontology, SEM, and the Polifonia CORE Ontology, with six purpose categories defined by domain experts: BusinessCareer, PersonalLife, Coincidence, Education, PublicCelebration, and MusicMaking.

The KE pipeline combines elements of traditional pipelines adapted to the MEETUPS scope:

- **Entity Recognition:** DBpedia Spotlight[15] identifies and links people and places to DBpedia resources, with quality control filtering to verify temporal consistency (ensuring participants could have actually met based on birth/death dates).

- **Temporal Processing:** A rule-based tagger based on SynTime [229] extracts temporal expressions, enhanced with NLTK[16] implementation and domain-specific heuristics. The system classifies expressions as time ranges, time points, or time references, achieving automatic normalization 65% to standardized data format (ISO8601), extended to 82% through LLM-assisted processing using GPT-3.5-Turbo for complex expressions.[17]

[15] `https://www.dbpedia-spotlight.org/`
[16] `https://www.nltk.org/`
[17] Morales Tirado et al. [137] refer to the model as ChatGPT without specifying the exact

- **Purpose Classification:** Initial machine learning classification achieved 46% precision, enhanced by zero-shot learning with GPT-3.5-Turbo, improving accuracy to 85% precision for the identification of meet-up purposes.

- **Coreference Resolution:** SpaCy's coreferee library [94][18] identifies implicit entity mentions (pronouns, noun phrases) to maximize entity detection coverage.

The harmonization algorithm processes extracted sentences to determine whether the text describes a meetup or not. Additionally, heuristic rules allow missing entities (time, place, or people) to inherit from adjacent sentences when they refer to the same event, enabling reconstruction of complete meet-up descriptions from fragmented textual evidence.

The evaluation methodology employs a dual approach combining Competency Question (CQ) testing and comprehensive domain expert validation. The first evaluation builds SPARQL queries mapped to CQs, where questions about a musician's life or meetups are evaluated. The second evaluation involved 12 domain experts validating the resource's value and utility. The expert panel comprised researchers (75%), musicians (33%), educators (33%), and historians (25%), with 91.7% reporting daily engagement with music-related content. Results demonstrated unanimous agreement (100%) on the value of documenting musical history encounters, with 50% rating it as "important" and 50% as "very important." Significantly, all respondents reported being unaware of any existing tool or database for storing and organizing historical music encounters.

The system is accessible to both scholars and developer communities through an open SPARQL endpoint, enabling integration with mapping li-

GPT model used in the article. The project's GitHub repository (`https://github.com/polifonia-project/meetups_pilot/`) reveals that the scripts use *gpt-3.5-turbo*, a legacy model (as of October 2025) estimated to be around 20B parameters, although OpenAI never disclosed the official size. The model card (`https://platform.openai.com/docs/models/gpt-3-5-turbo`) provides additional technical details.

[18] `https://spacy.io/universe/project/coreferee`

braries, GIS software, timeline displays, and educational applications. The resource follows FAIR principles with availability through GitHub, permanent SPARQL endpoint access, and related documentation.

MusicBO MusicBO[19] represents a specialized text-to-KG approach from domain-specific musical heritage documentation, focusing specifically on the music history of Bologna [70]. The project addresses the challenge of processing heterogeneous historical documents containing long-tail entities that are not easily reconcilable with general-purpose knowledge bases, demonstrating how text-to-KG pipelines can be adapted for highly specialized CH corpora. The project processed a carefully curated multilingual corpus of 137 texts spanning from the 1700s to the present day, including 47 English documents, 51 Italian documents, and materials in French and Spanish. The corpus encompasses diverse textual genres relevant to Bologna's musical heritage: correspondence between musicians and cultural figures, biographical narratives, and musicological studies. Table 8.14 summarizes key characteristics of MusicBO. The approach used in MusicBO differed significantly from MMKG. While the latter, starting from Wikipedia, has an inherent advantage (e.g., DBpedia Spotlight, sitelinks), MusicBO's corpus represents a domain-specific collection with long-tailed entities that are not easily reconcilable to DBpedia or Wikidata. The project implements a pipeline combining Text2AMR2FRED[20]with automated quality control measures. The process transforms sentences through neural models (SPRING[21] for English) to create AMR graphs, which are then converted into RDF/OWL KGs. The resulting graphs are enriched using Framester[22] alignments with external knowledge bases, including DBpedia, Wikidata, and VerbAtlas.[23] The evaluation approach addresses the challenge of assessing text-to-KG quality for historical texts through back-translation. AMR graphs undergo quality assessment by converting them back to natu-

[19] https://github.com/polifonia-project/musicbo-knowledge-graph/
[20] https://framester.istc.cnr.it/txt-amr-fred/api/docs
[21] http://nlp.uniroma1.it/spring/
[22] https://framester.github.io/
[23] https://verbatlas.org/

ral language and comparing with original sentences using BLEURT scores for English and cosine similarity for Italian. Graphs paired with sentences scoring below established thresholds (negative BLEURT scores or cosine similarity below 0.90) are filtered out. The pipeline initially processed 62,377 sentence-AMR pairs but retained only 7,557 pairs after quality filtering, demonstrating the importance of rigorous validation for historical documents. These filtered pairs generated 531,073 triples in the final KG, with English documents contributing 412,911 triples and Italian documents contributing 118,162 triples. The resulting KG enables large-scale qualitative and quantitative analysis of Bologna's musical heritage corpus and supports the creation of visual data stories through MELODY. The system successfully captures complex musical events and relationships, such as engagement offers between opera managers and composers, with entity linking to Wikipedia, Wikidata, and DBpedia whenever available.

Viewsari

Viewsari is an ongoing project[24] that focuses on creating a KG of Giorgio Vasari's seminal work *Lives of the Most Eminent Painters, Sculptors, and Architects* (often referred to as *The Lives*), a Renaissance text published in 1550 and expanded in 1568 [141]. This project—similarly to the *Vespasiano da Bisticci. Lettere* project discussed in Chapter 1—uses text-to-RDF to enrich the navigation of *The Lives* and support qualitative and quantitative analysis of Vasari's descriptions. Table 8.15 summarizes key characteristics of Viewsari. The ontological model behind Viewsari implements eXtreme Design, a user-centered design approach that emphasizes stakeholder involvement throughout the ontology engineering process. Art historians provided insight into the specific requirements for representing Vasari's Renaissance world [141]. The project introduces a dedicated `:Cooccurrence` class that connects a minimum of two entities (e.g., an artist and a work) and is annotated with provenance information linking it to specific paragraphs in the text. The text-to-KG pipeline

[24] `https://viewsari.ise.fiz-karlsruhe.de/`

processes the English translation by Gaston C. Du Vere (1912), and implements a traditional architecture adapted to satisfy the ontological model in use [172]. The NER step identifies various entity types with different degrees of granularity: Person, Locations, Artifact, Material, Technique, Motifs, and Dates. Entity linking to existing authority files like VIAF[25] and Wikidata presents challenges, particularly for artworks, materials, and places that may have changed names over time or represent long-tail entities not easily reconcilable with general-purpose entity linking tools. The project also addresses the specialized challenge of linking art motifs to the ICONCLASS[26] classification system. The social network extraction component employs Pointwise Mutual Information (PMI) to quantify the strength of relationships between artists based on their co-occurrence patterns. The networks across all 10 volumes of Vasari's work demonstrate clear patterns, with high PMI values often corresponding to documented artistic collaborations such as those between Piero di Cosimo and Cosimo Rosselli, or Leonardo da Vinci and Verrocchio. The KG has two layers: one referring to the content of the work, one describing its structure. The two layers are connected through the Provenance-Content pattern, representing source-evidence for extracted entities and co-occurrences. This pattern enables precise tracing of assertions back to specific passages in Vasari's text, maintaining scholarly rigor essential for art-historical research. The system provides interactive network visualizations accessible online and converts the social network into a KG format with SPARQL query capabilities, facilitating integration with broader semantic web resources. As an ongoing project, exact information on the size of the resulting KG is not yet available.

Bernoulli-Euler-Digital

The Bernoulli-Euler Digital project[27] exemplifies KE from historical scientific texts as part of the larger Bernoulli-Euler Online (BEOL) initiative funded by the Swiss National Science Foundation [7]. This project integrates previously

[25] `https://viaf.org/en`
[26] `https://iconclass.org/`
[27] `https://bernoulli-euler.dhlab.unibas.ch/`

incompatible edition projects—including the *Basler Edition der Bernoulli-Briefwechsel* (BEBB), *Leonhardi Euleri Opera Omnia* (LEOO), and Jacob Bernoulli's scientific notebook *Meditationes*—into a unified platform. Table 8.13 summarizes key characteristics of Viewsari. To extract information about the subjects' movements and travels, the project implements the "textToRDFGraph" pipeline [7]. The workflow operates in two stages: NER identifies relevant entities, followed by supervised relation extraction based on ontological constraints to detect trips and movements. The pipeline employs RDF-star to link extracted assertions directly to their source documents within the digital edition. For instance, the assertion that Euler lived in Berlin can be represented as `«:euler :livedIn :berlin» :mentionedIn :document1`, where the nested triple structure preserves the provenance of the extracted information [7]. The resulting KG enriches entity descriptions with external information from Wikidata and supports network visualization as interactive 3D force-directed graphs, enabling researchers to explore correspondence networks, including connections to the scientific letters of Isaac Newton and Leibniz. The project also integrates TEI/XML[28] encoding and LaTeX representation of mathematical notation to properly model the content of the documents [7].

2.3.3 Specialized Documents

InTaVia

The InTaVia (In/Tangible European Heritage: Visual Analysis, Curation & Communication) project integrates biographical and CH data from different sources in a single KG [181]. As part of an H2020 research initiative, InTaVia bridges the gap between tangible cultural artifacts and intangible biographical information by creating a transnational KG that connects previously isolated data collections across Europe. Table 8.10 summarizes key characteristics of InTaVia. The KG integrates data from four national biographical

[28] `https://www.tei-c.org/`

dictionaries: Austria (APIS),[29] Finland (BiographySampo), Slovenia (SBI),[30] and the Netherlands (BiographyNet).[31] These biographical data are harmonized according to the InTaVia Data Model (IDM-RDF) [114], which is based on CIDOC CRM. After the integration and harmonization of all biographical datasets, related cultural objects were retrieved from Europeana and Wikidata for each actor. The resulting InTaVia KG contains 24,588,310 triples on 165,960 actors from the four participating countries, 230,068 cultural objects from Europeana, 160,239 from Wikidata, as well as 3,257 institutions and 24,446 places [114]. The system processes information about 112,050 persons described by 257,673 person proxies. The majority of documented biographical events concentrate in the 19th and 20th centuries. InTaVia applies text-to-KG on Wikipedia biographies to enrich its data. The pipeline employs a dual-framework approach with support for multiple languages (English, Dutch, Finnish, and Slovenian). The Flair pipeline [191] performs preprocessing, NER, semantic frame disambiguation, relationship extraction, and entity linking. The complementary AllenNLP pipeline [74] specializes in proposition-level and discourse-level annotation, covering Semantic Role Labeling and coreference resolution. The evaluation was carried out through a comparative analysis of the generated RDF, comparing quantitative and qualitative differences from DBpedia and Wikidata. Their "closer reading" methodology revealed that the pipeline could capture new biographical information not present in existing resources [187]. The InTaVia platform[32] consists of three main modules: the Data Curation Lab for searching and curating data, the Visual Analytics Studio for analyzing data through multiple visualization types (maps, timelines, network graphs), and the Storytelling Suite for communicating cultural information to non-expert audiences. The platform supports "post-anthropocentric storytelling," which enables different types of entity to serve as protagonists in cultural narratives.

[29] https://github.com/acdh-oeaw/apis-core
[30] https://www.obrazislovenskihpokrajin.si/en/
[31] https://research.vu.nl/en/projects/biographynet
[32] https://intavia.eu/

Notarypedia

Notarypedia is a project that addresses the challenge of making the registers of the La Valletta Notarial Archives queryable through entities and relations. It represents a specialized approach to extracting knowledge from historical legal documents, focusing on 15th-century Maltese Notarial Acts housed in the Notarial Archives in Valletta [60]. Notarypedia addresses the challenge of making over 20,000 historical manuscripts from the Maltese Notarial Archives accessible to diverse users, including notaries, historical researchers, and genealogists [60]. Table 8.12 summarizes key characteristics of Notarypedia. The corpus consists of 15th-century registers written primarily in Latin with medieval Sicilian and Maltese elements, presenting unique linguistic challenges that reflect the cultural complexity of medieval Malta. The project implements a pipeline that adds a reasoner to the traditional pipeline architecture. Entities are recognized using NER models trained on publication indices to identify persons, places, dates, and domain-specific keywords, complemented by rule-based temporal extraction that handles the indiction date methodology common in historical documents. Coreference resolution handles frequent name variants within documents using multiple similarity measures, with rule-based approaches achieving optimal performance (F_1 score of 0.96). Relation extraction focuses on 14 types of relationships that span genealogical, geographical, and commercial connections. The system employs supervised machine learning to identify relationships between pairs of entities, with logistic regression demonstrating the strongest performance (70% accuracy, 68% F_1 score) for extracting relationships such as family relationships and spatial connections. The KGs are then completed through two complementary mechanisms: logical inference using Apache Jena's rule reasoner to deduce additional genealogical relationships through forward chaining, and KG embedding models for link prediction, with TransE [119] achieving 49% accuracy for predicting missing relationships. The resulting KG employs a Notarial Ontology that extends

FOAF,[33] Schema.org,[34] and Getty Vocabularies[35] to capture the specialized semantics of medieval commercial and legal relationships [60].

2.3.4 Comparative Analysis Across Projects

Table 2.3 provides a systematic comparison of the 11 projects analyzed in the eight analytical dimensions established in the methodology.

Table 2.3: Comparative Analysis of Text-to-KG Projects

Project	**Primary Method**	**Core Ontology**	**Evaluation Approach**	**KG Size (triples)**
BiographySampo	Pattern + NLP	Bio CRM	Manual (135 events)	10M
ParliamentSampo	Rule-based	Custom + CIDOC CRM	Application-based	Not reported
WarSampo	AATOS annotation	CIDOC CRM	P/R (433 articles)	14M
HyperReal	Pattern matching	Simulation Ontology	F_1 score	500K
Odeuropa	BERT multitask	CIDOC CRM + CRMsci	F_1 per frame element	2.4M instances
MMKG	DBpedia + LLM	MEETUPS + PROV	CQ + Expert validation	45,812 meetups
MusicBO	AMR + FRED	Framester alignment	Back-translation	531K
Viewsari	Traditional NLP	eXtreme Design	PMI network analysis	Not reported
Bernoulli-Euler	Supervised RE	RDF-star	Not reported	Not reported
InTaVia	Flair + AllenNLP	IDM-RDF (CIDOC CRM)	Closer reading vs. DBpedia	24.6M
Notarypedia	ML + Reasoner	Notarial Ontology	F_1 + Link prediction	Not reported

Several patterns emerge from this comparative analysis. Rule-based and pattern-driven approaches, exemplified by the Sampo projects and HyperReal, demonstrate robust performance on semi-structured sources, achieving precision rates between 82-99% on domain-specific tasks. This performance level is expected given the formulaic nature and the pre-existing structure of their input documents. Machine learning and deep learning approaches, represented by projects such as InTaVia and Odeuropa, show competitive performance on previously underrepresented extraction tasks. Odeuropa achieves F_1 scores of 65.5-88.7% for core olfactory frame elements, representing the first systematic attempt at this specialized extraction challenge. Variations of the traditional

[33] `http://xmlns.com/foaf/spec/`
[34] `https://schema.org/`
[35] `https://www.getty.edu/research/tools/vocabularies/`

pipeline architecture presented in Section 2.6.1 constitute the most common approach (7 of 11 projects). Notable variations include: harmonization algorithms and LLM-enhanced processing in MMKG for temporal normalization and purpose classification; Notarypedia's reasoner component for link prediction; provenance modeling in Viewsari and Bernoulli-Euler Digital. All projects employ closed KE approaches in their pipeline design, with MusicBO as the sole exception through its use of text2AMR2FRED, a tool we discussed in Section 1.2.3.

Evaluation Practices Evaluation approaches demonstrate a variety closely related to the motivation of the project and the intended audience. Projects serving domain experts employ domain expert validation (MMKG, Viewsari), competency question testing (MMKG, HyperReal), and comparative analysis against established resources (InTaVia) as additional validation mechanisms. Precision, recall, and F_1 scores constitute the most common quantitative metrics, calculated either automatically or through human annotation. This heterogeneity reflects the absence of standard evaluation protocols for KG generation, where human evaluation remains the gold standard [132].

Bottlenecks and Persistent Challenges The analysis reveals bottlenecks stemming from two primary sources: document characteristics and pipeline component performance. Document-level challenges include degradation of OCR quality (WarSampo, MusicBO, Bernoulli-Euler), historical language variation (Notarypedia, Bernoulli-Euler), and semantic change between temporal contexts (Odeuropa). Component-level bottlenecks focus on NER, relation extraction, and, particularly, entity linking, which emerges as the challenge most frequently addressed across projects. Entity type coverage demonstrates an unexpected pattern: most projects extract common entity types (person, location, organization) and relations, enabling reuse of existing NER tools such as Flair and SpaCy rather than requiring domain-specific training. Odeuropa represents a notable exception, necessitating specialized models for olfactory frame elements. Vocabulary alignment rarely constitutes a significant challenge

due to the closed KE architecture's vocabulary-guided design, except in multilingual contexts (InTaVia, MusicBO). Specialized evaluation methodologies emerge as necessary complements to traditional NLP benchmarks, which fail to capture humanities scholarship requirements. Back-translation validation (MusicBO) and combined competency question testing with domain expert validation (MMKG) represent the most transferable approaches beyond standard precision, recall, and F_1 metrics.

Project Motivations The analyzed projects reveal two distinct motivations for applying text-to-KG methodologies in CH contexts:

1. **Enhancement of Scholarly Digital Edition Navigation:** Projects like Viewsari and Bernoulli-Euler Digital apply text-to-KG to enable semantic navigation within digital editions. This approach transforms traditional reading interfaces into queryable knowledge structures, allowing researchers to trace relationships between entities across the text. This aligns with the methodology employed in the *Vespasiano da Bisticci. Lettere* project discussed in Chapter 1, where extracted KGs enrich linear textual access with network-based exploration capabilities.

2. **Computational Infrastructure for Research:** Projects including Sampo (BiographySampo, ParliamentSampo, WarSampo), Notarypedia, Odeuropa, HyperReal, the Polifonia pilots (MMKG, MusicBO), and InTaVia apply text-to-KG to transform textual sources into queryable knowledge bases. These projects enable computational approaches to research questions—whether prosopographical networks, olfactory heritage patterns, musical encounters, or cross-national biographical comparisons—that would be impractical through traditional close reading methods. The motivation is to create research infrastructure that supports both exploratory discovery and hypothesis validation at scale.

2.4 Challenges

The comparative analysis in Section 2.3 identified persistent bottlenecks across text-to-KG implementations. This section examines these challenges systematically, distinguishing between cross-domain problems affecting text-to-KG extraction broadly and domain-specific obstacles particularly acute in CH contexts. Rodriguez et al. [95] identify fundamental problems of ambiguity, alignment, and data sparsity affecting text-to-KG systems across domains. CH applications face these problems with particular intensity due to compounding factors: historical language variation, narrative discourse structures that embed factual claims within interpretive commentary, multilingual document corpora, and prevalence of long-tail entities absent from general-purpose knowledge bases. The following subsections examine these challenges through evidence from the analyzed case studies, organizing them into four categories that correspond to the pipeline bottlenecks identified in the comparative analysis.

2.4.1 Data Alignment Challenges

Regino et al. [163] characterize alignment as a primary challenge in Semantic Web applications, distinguishing between problems in *generating new triples* from text (language ambiguity, contextual dependencies, NER errors) and challenges in *integrating new triples* into existing KGs. Their framework identifies three critical alignment dimensions:

- **T-Box alignment:** New RDF triples may contain entity and property types undefined in the target ontology, requiring schema extension or mapping to existing concepts

- **A-Box alignment:** New triples may duplicate information already present in the KG, necessitating deduplication and consistency checking

- **URI alignment:** New resource identifiers must be validated against

existing entities to prevent duplicate resource creation for identical real-world entities

T-Box Alignment in Practice Projects address T-Box alignment through three primary strategies. InTaVia [114] developed the IDM-RDF ontology to harmonize four national biographical dictionaries, employing the OAI-ORE model to preserve institutional perspectives while mapping to common CIDOC CRM concepts. Essentially, multiple proxies are used to get from the extracted terms to the ontology and vocabularies. MusicBO [70] achieved alignment through Framester mappings to external knowledge bases (DBpedia, Wikidata, VerbAtlas), enabling domain-specific AMR-derived properties to align with established vocabularies. This out-of-the-box alignment bypasses any additional alignment with other ontologies. However, queryability is affected as linguistic mappings are heavily impacted by the structure of the input sentence. Odeuropa [122] extended CIDOC CRM and CRMsci with new classes and properties for olfactory information, ensuring alignment with the annotation by design. Other projects which did not account for alignment by design had to sort to additional correction (through rules or by hand).

A-Box Alignment and Deduplication A-Box alignment solutions vary based on data source characteristics. InTaVia addressed deduplication through sameAs linking and cross-dataset reconciliation, though only 548 cross-dataset links were established across four national collections, indicating persistent challenges in entity resolution at scale. WarSampo [110] implemented quality control filtering and temporal consistency validation to prevent duplicate or temporally inconsistent information. BiographySampo [115] employed name and birth year comparison for external source linking across 16 databases with systematic deduplication procedures.

URI Alignment and Entity Linking URI alignment approaches demonstrate tension between automated and manual strategies, corresponding to the entity linking bottleneck identified in the comparative analysis. MMKG [137] employed DBpedia Spotlight with quality control filtering based on tempo-

ral consistency (ensuring participants could have met based on lifespan dates). Viewsari [114] combined the Index of Names with VIAF and Wikidata linking, though long-tail entities (Renaissance artworks, materials, iconographic motifs) presented significant challenges that, as an ongoing project, have not yet been explicitly addressed. Bernoulli-Euler [6] implemented authority control through GND numbers for people and Geoname IDs for locations. HyperReal [174] used owl:sameAs linking while preserving context-dependent symbolic meanings across cultural contexts. The prevalence of long-tail entities in CH corpora — artworks known only to specialists, local historical figures, domain-specific terminology — limits the effectiveness of general-purpose entity linking tools trained primarily on Wikipedia and news corpora. This necessitates project-specific entity indexes (e.g. Viewsari's Index of Names) or acceptance of lower linking rates with manual curation for critical entities.

Coreference Resolution and Entity Disambiguation Coreference resolution presents acute challenges in CH contexts where naming conventions are inconsistent and contextual information is sparse [59, 108]. The same full name can refer to different persons, and different names can refer to the same person within genealogical networks. The extraction process of BiographySampo, for instance, revealed that four people named Christian Trapp (grandfather, father, son, grandson) could be distinguished only by lifespan [115]. WarSampo [110] demonstrates how data sparsity compounds extraction errors: military ranks change over time, names may be spelled differently or incompletely recorded, and copy errors in source materials or OCR processes exacerbate disambiguation problems. These challenges exceed the capabilities of contemporary coreference resolution systems trained on modern, well-edited texts with consistent naming practices. However, coreference resolution constitutes a bottleneck only when the KG scope requires entity resolution across discourse boundaries. Projects with sentence-level or event-level extraction scope avoid this challenge entirely. In the case of Odeuropa, each smell event KG corresponds to a single sentence/paragraph or minimal discourse unit,

with entity references resolved within that local context. The project does not require identifying that "the perfume" in paragraph three refers to "jasmine essence" mentioned two pages earlier, as each extracted event stands independently.

2.4.2 Language Processing Challenges

Multilingual and Historical Language Variation CH texts present linguistic challenges that extend beyond the contemporary multilingual NLP. Notarypedia [60] processes 15th-century Maltese Notarial Acts written primarily in Latin with medieval Sicilian and Maltese elements. Bernoulli-Euler Digital [6] includes historical letters in German, French, and Latin that exhibit linguistic variation between single authors over time. Odeuropa [122] addresses multilingual challenges across six European languages, developing frame-based annotation schemes for olfactory experiences. Frame recognition performance varies significantly between languages, with English dominating training data. Historical language change and polysemous smell-related words (terms with different meanings across temporal and cultural contexts) introduce additional complexity, thus contextualizing the annotation and custom training for the text-to-KG approach they used. These challenges correspond to the document-level bottleneck identified in the comparative analysis: historical language variation degrades NER performance, as models trained on contemporary language struggle with orthographic variation, obsolete terminology, and semantic shift.

Temporal Processing and Normalization The effectiveness of temporal processing depends on the origin, composition date, and domain conventions of the document. Notarypedia [60] employs rule-based approaches for the indiction date methodology common in notarial documents, where traditional temporal extraction tools fail due to non-standard calendrical systems and the large historical time span (five centuries). MMKG [137] demonstrates the complexity of temporal processing, achieving automatic normalization 65% of dates to a standard format, extended to 82% through LLM-assisted processing

with GPT-3.5-Turbo for complex expressions. Even with LLM enhancement, temporal extraction, even on "standard" documents like Wikipedia, requires manual intervention and domain-specific approaches. BiographySampo [115] achieves 98% accuracy through statistical reasoning about historical lifespans and family relationships, demonstrating that domain knowledge applied to pattern-based extraction can surpass deep learning approaches for structured temporal information. However, this success is limited to biographical texts with formulaic temporal expressions.

2.4.3 Entity Recognition and Specialized Vocabularies

Long-tail Entity Recognition CH resources contain specialized entity types that rarely appear in general corpora, creating recognition and linking challenges identified as the primary bottleneck in comparative analysis. Viewsari relies on *The Lives*' internal Index of Names on top of general-purpose authority files. MusicBO mints new entities for Bologna's musical heritage, linking only those present in Wikidata, Wikipedia, or DBpedia — a minority of mentioned entities. Other projects (Notarypedia, Bernoulli-Euler) rely on manual alignment for entities absent from standard knowledge bases. The comparative analysis revealed that most projects extract common entity types (person, location, organization), enabling reuse of existing NER tools (Flair, SpaCy) without domain-specific training. This suggests that entity type recognition is not the bottleneck; rather, entity linking for recognized entities constitutes the challenge when entities fall outside Wikipedia-derived knowledge base coverage.

Domain-Specific Terminology and Novel Entity Types Projects which approach new terms and novel entity types demonstrate that low performance on novel extraction tasks necessitates either substantial annotation effort for supervised training or acceptance of manual intervention for quality control. Odeuropa [122] pioneered olfactory experience extraction, developing a specialized frame-based annotation scheme with nine Frame Elements. HyperReal [174] presented a KG of symbolic meaning with a KE pipeline. Entirely new

CH information categories can be systematically extracted when appropriate frameworks are developed, though requiring substantial manual annotation effort. Viewsari, apart from typical types, needs NER to recognize artworks and other works, similarly to Notarypedia. The first leveraged spacy, while the second used extensive lexicons to detect products being sold.

2.4.4 Document Quality and Evaluation

OCR Quality and Preprocessing OCR quality impacts downstream extraction performance, corresponding to the document-level bottleneck in the comparative analysis. WarSampo [110] reports precision improvements between 75.76%-82.02% for military unit identification after specialized OCR post-processing. Bernoulli-Euler Digital, Odeuropa, and MusicBO required careful document correction to avoid errors. No domain-agnostic OCR correction solution emerged across projects, with each implementing corpus-specific error correction rules.

This suggests that OCR quality control remains a manual, resource-intensive process that requires domain expertise to identify and correct errors that propagate through extraction pipelines.

Evaluation Methodologies The comparative analysis identified evaluation heterogeneity reflecting the absence of standard KG generation protocols. F_1, precision, and recall constitute common quantitative metrics, though calculation methods vary (automatic vs. human annotation). Projects serving domain experts supplement quantitative metrics with qualitative validation approaches. MusicBO [70] implements back-translation validation using BLEURT scores for English and cosine similarity for Italian, filtering graphs scoring below established thresholds. This rigorous filtering retained only 7,557 of the 62,377 sentences-AMR pairs initially processed, demonstrating that quality control can dramatically reduce output volume. MMKG used competency question testing and domain expert validation (12 experts), establishing resource utility through user studies rather than extraction accuracy metrics alone. These approaches suggest that evaluation in CH contexts

requires domain-specific validation beyond standard NLP benchmarks, which measure performance on tasks (NER, relation extraction) rather than end-user utility for humanities research questions. However, both could have been implemented across most of the projects of this survey, as the minimal requirements are common to each of these projects.

2.5 Trends

The second literature analysis of this work identified methodological trends, starting from a list of keywords that we compiled from the challenges documented in Section 2.4. However, part of these trends do not directly map to the identified challenges. Of 227 papers published between 2020-2025 retrieved through the methodology described in Section 2.2, quantitative analysis reveals concentration in three technological approaches: 44 papers mention LLMs in titles, 11 address KG embeddings, and 9 focus on Graph Neural Networks. Publication patterns show acceleration from 2023-2025, with 2024 as the most prolific year (98 papers, including 8 explicitly addressing CH domains). In the rest of the paper, we will illustrate some of the most interesting works of the selections. Papers primarily address three pipeline subtasks: KG construction (55 papers), Relation Extraction (35 papers), and NER (31 papers). The domain distribution shows that 150 papers target general-purpose applications, while the remainder focus on CH and adjacent fields (educational, literary, and historical contexts). This temporal pattern aligns with the surveyed projects' timeline: text-to-KG scalability improved first through Transformer architectures, then through LLMs. The following subsections examine trends addressing specific challenges identified in Section 2.4, organized by pipeline bottlenecks they target: entity recognition challenges (Section 2.5.1), relation extraction bottlenecks (Section 2.5.2), entity linking for long-tail entities (Section 2.5.3), and evaluation methodologies (Section 2.5.4). In particular, the literature provides limited coverage of certain challenges identified in the case studies, such as co-reference resolution in CH-specific contexts. The selection

of articles for detailed analysis was based on two criteria: demonstrated application to CH domain contexts and implementation of KG generation methodologies that can be adapted to different project requirements.

2.5.1 Named Entity Recognition for Specialized Contexts

NER research addresses the entity recognition bottleneck identified in Section 2.4.3, particularly the challenge of domain-specific terminology and historical language variation discussed in Section 2.4.2. Recent developments focus on enhancing performance for fine-grained entity types beyond traditional Person, Location, and Organization categories while tackling domain-specific obstacles in CH contexts. The scarcity of specialized training data remains a persistent challenge despite advances in transfer learning and pre-trained models. Two complementary approaches have emerged to address this limitation. The first involves automatic dataset generation techniques. Jain et al. [104] recognize that creating NER datasets for each specific task may be infeasible, proposing instead a three-stage framework that leverages existing CH resources from knowledge bases such as Wikidata and Getty vocabularies to automatically generate high-quality training data for identifying artwork titles in digitized art archives. This approach demonstrates how domain-specific NER challenges can be addressed through innovative data generation strategies combining dictionary-based matching, quality filtering with weakly supervised learning systems such as Snorkel, and enhancement with silver standard annotations from Wikipedia. The second approach involves generalist NER models that adapt to new contexts without extensive retraining such as GLiNER (Generalist and Lightweight Named Entity Recognition), which represents a significant development in this direction [225]. Unlike traditional NER models trained on fixed entity types, GLiNER recognizes arbitrary entity types specified through natural language descriptions, making it valuable for specialized domains where annotated data is scarce. The temporal dimension presents challenges particularly acute for historical documents, as

discussed in Section 2.4.2. Language, entity, and concept drift reduce NER quality, as tools trained on modern language may not transfer effectively to historical variants. Ehrmann et al. [57] identify this challenge of "dynamics of language" as manifesting through systematic spelling variations from orthographic reforms, evolving naming conventions with obsolete titles and address forms, and entity drift where professions and entity types change over time. These temporal variations create sparser feature spaces for NER systems, with studies showing F_1-score drops from 87% to 63% for person entities when encountering poor OCR quality combined with historical language variations. Researchers have responded by developing temporal-aware historical NER approaches accounting for evolution of entity naming conventions and historical context [79]. Pre-trained language models (PTLMs) have become essential tools for addressing these temporal challenges. Santini et al. [171] demonstrate this by applying PTLM models to Giacomo Leopardi's Zibaldone, a 19th-century Italian scholarly text. Their experiments, conducted on a dataset containing 2,899 references to people, locations, and literary works, reveal that while instruction-tuned LLMs encounter difficulties with historical humanistic texts, fine-tuned NER models offer more robust performance even with challenging entity types such as bibliographic references. Contemporary research has also focused on addressing bias and representation issues in NER for CH, particularly in colonial archives where traditional approaches may fail to capture voices of marginalized communities [130]. These efforts highlight the importance of developing inclusive NER approaches that can surface previously hidden or underrepresented entities in historical records.

2.5.2 Relation Extraction Advances

Recent advances in relation extraction address the bottleneck identified in Section 2.4, expanding from traditional binary relations to more complex graph structures required by CH applications. RE datasets have become more available and fine-grained in supervised settings. Ali et al. [9] created one of the most extensive joint NER and RE resources with RELD, containing almost

1,230 million triples describing 1,034 relations, 2 million sentences, 3 million abstracts, and 4,013 documents in RDF format (NIF). Wikidata has become an essential resource for generating specific datasets: while its data model is less formalized than generalist ontologies such as Schema.org, the number of relationships and volume of data make it a cornerstone for reuse. Diverse tools approach RE not as multi-class classification but as a generative task [96]. This enables more generalist tools such as REBEL[36] or, for joint NER, RE, and EL, tools such as RELIK[37] [145], which largely extend the range of possible relations to be extracted. For more complex graph shapes including qualifiers and *n*-ary relations, automatic approaches have emerged, such as the HyperRED dataset [35] and attempts to automatically extract n-ary relationships [128]. However, in CH contexts, most approaches rely on pipelines to overcome the limitation of extracting plain relations [174, 179]. BERT can also be used as a Semantic Role annotator aligned to map to CIDOC CRM graphs [206]. For unsupervised RE, LLMs have become state-of-the-art, capable of automatically inferring relations to focus on from text and adaptable through context engineering (especially few-shot learning) to produce desired outputs [116]. Recent LLMs with larger context windows can perform RE on documents rather than only paragraphs or simple texts, addressing coreference resolution issues [220]. LLMs have been used efficiently in CH RE contexts. Santini [170] demonstrates performance exceeding mREBEL on 1800s Italian texts. Giagnolini et al. [78] present a proof of concept using LLMs for text-to-KG of archival string fields. A pipeline based on LLama3.3-70B processes the biography of Andrea Costa, the archive originator, by first classifying each paragraph according to relevant event classes, then applying specific schemas depending on paragraph classification (birth event, death event, political event). The relations are then assigned to the RiC-O ontology.[38] Beyond linking entities to external KBs, CH projects often require matching entities across multiple internal datasets,

[36] `https://github.com/Babelscape/rebel`
[37] `https://github.com/SapienzaNLP/relik`
[38] `https://www.ica.org/resource/records-in-contexts-ontology/`

a task known as entity matching or entity resolution. Baas et al. [13] address this challenge in decentralized CH data contexts where multiple datasets contain duplicates within and between sources. Their approach handles complex properties characteristic of historical data: missing attributes, approximate dates, name variants (e.g., "Jans," "Jansz," "Janssen" referring to the same person), and proxy variables where different datasets use related but distinct attributes (birth dates vs. baptism dates).

2.5.3 Addressing the Long-Tail Entity Linking Challenge

EL research directly addresses the primary bottleneck identified in comparative analysis (Section 2.3): linking entity mentions to specific identifiers in knowledge bases for disambiguation and KB enrichment, as discussed in Section 2.4.3. Recent empirical research has quantified the severity of the long-tail problem in CH contexts. The KE-MHISTO benchmark demonstrates that historical texts contain significantly higher concentrations of long-tail entities compared to mainstream EL datasets [81]. While traditional benchmarks such as AIDA CoNLL-YAGO focus on highly popular entities, historical music periodicals show 30% in English and 28% in Italian datasets lacking corresponding Wikidata entries (NIL entities). This distribution has profound performance implications: state-of-the-art models achieve F_1 scores of only 0.47-0.61 on historical texts compared to 0.88-0.90 on contemporary datasets. Fine-grained entity linking work by Rosales-Méndez et al. [166] reveals that different EL systems target fundamentally different entity types, with some focusing exclusively on named entities while others include common entities. This variability becomes particularly problematic in CH domains where specialized vocabularies and entity types are prevalent. The KG-ZESHEL approach by Ristoski et al. [165] addresses the challenge of linking mentions to unseen entities without extensive labeled data. By improving BERT-based EL with auxiliary information from KGs—entity types, popularity scores, and graph embeddings—the system achieves improvements in candidate generation and ranking tasks. This approach is particularly relevant for CH applications where new entities are fre-

quently encountered but where existing KBs are also present (e.g., Viewsari's Index of Names discussed in Section 2.3.2). For CH contexts, the ability to handle NIL (unlinkable) entities becomes crucial. Evaluation on KE-MHISTO shows that LLMs outperform specialized EL models when explicitly instructed to predict NIL entities, with LLAMA 3.3 70B achieving F_1 scores of 0.48-0.51 compared to 0.37-0.47 for traditional systems such as mGENRE and BELA [81]. Recent research has exposed significant gender biases in both CH texts and knowledge bases. Historical music corpora show male entities account for 75-85% of person mentions, with female entities being disproportionately classified as NIL (50-55% versus 27-28% for males) [81]. These biases compound the long-tail challenge, as female historical figures are both less frequently mentioned and less likely to have Wikipedia coverage. Fine-grained entity linking work proposes configurable performance measures based on fuzzy sets adaptable for different application scenarios [166], recognizing that different entity types and links may have varying importance depending on specific CH applications. Emerging approaches show promise for CH entity linking. Using GLiNER with an additional lexicon achieves competitive performance (F_1 scores of 0.59-0.63) compared to much larger language models while allowing recognition of arbitrary entity types specified through natural language descriptions [81]. Multilingual models demonstrate particular potential, with smaller multilingual systems achieving performance comparable to significantly larger counterparts for CH applications.

2.5.4 Evaluation Methodologies Beyond Traditional Metrics

The increasing complexity of automatically generated KGs has driven the development of sophisticated evaluation methodologies that go beyond traditional precision-recall metrics, addressing the evaluation heterogeneity identified in Section 2.3. Two complementary trends have emerged: LLM-based evaluation frameworks that assess semantic quality and domain-specific adequacy of extracted knowledge structures, and perspectivist approaches that challenge

the assumption of single ground truth annotations. Traditional evaluation frameworks in NLP have operated under the assumption that disagreement among annotators represents noise requiring resolution into a single ground truth through aggregation, adjudication, or statistical methods. Perspectivist NLP challenges this assumption, recognizing that multiple valid perspectives may coexist for subjective interpretation tasks [29]. This approach proves particularly relevant for CH contexts where scholars may legitimately disagree on entity boundaries, relationship interpretations, or event classifications based on different theoretical frameworks or domain expertise. Strong perspectivism extends beyond data collection to encompass modeling, evaluation, and explanation [29], aligning with principles established in Data Statements [20] and Data Envelopes [129] for transparent documentation of dataset characteristics. This parallels broader trends in CH knowledge representation [147] and extraction [179] that recognize concurrent opinions as integral to CH dataset collection rather than as annotation artifacts requiring resolution. Traditional KG evaluation also faces fundamental challenges when structural alignment between generated and reference graphs is impossible due to different modeling approaches or entity resolution strategies [73, 70]. MusicBO applied back-translation validation in CH contexts using BLEURT scores for English and cosine similarity for Italian, filtering out graphs scoring below established thresholds [70]. This methodology enables comparison when test data is not available, using metrics such as BLEU [146]. METEOR [15] extends beyond BLEU's n-gram matching by incorporating stemming, synonymy, and paraphrase detection, making it suitable for evaluating semantic equivalence in KG content. BARTScore [224] leverages pre-trained BART models to provide contextual similarity assessment capturing semantic relationships rather than surface-level similarity, particularly valuable for KGs where vocabulary might have high mismatches with input text. CHRF++ [157] focuses on character-level n-gram matching combined with word-level features, proving robust for evaluating multilingual content. G-EVAL [123] fully implements the LLM-as-

a-judge approach [91]. Unlike similarity-based metrics that can be biased when comparing source documents with retranslated versions of KGs representing only subsets of original information (for instance, in closed KE), G-EVAL offers flexibility through tailored instructions specific to domains, for instance by combining them with competency questions.

2.5.5 LLMs in Ontology Engineering and Text-to-KG

The publication acceleration observed in the literature analysis (44 papers mentioning LLMs in titles from 2020-2025, with concentration in 2023-2025) reflects broader developments in applying LLMs to ontology engineering and KG construction tasks. While the surveyed CH projects primarily employ rule-based, pattern-based, or supervised learning approaches, recent research investigates LLMs as components in text-to-KG pipelines and ontology development workflows, raising questions about their potential role in addressing CH-specific challenges identified in Section 2.4.

Garijo et al. [75] provide a landscape analysis mapping LLM applications to ontology engineering tasks using the Linked Open Terms (LOT) methodology. Their categorization reveals that existing efforts concentrate on early-stage activities: competency question generation, term typing, taxonomy discovery, and ontology encoding from natural language descriptions. Notable gaps remain in ontology evaluation, documentation, and maintenance tasks. The analysis identifies three principal challenges for LLM-based ontology engineering: establishing common task definitions with standardized inputs and outputs, developing evaluation methods that accommodate multiple valid ontological representations rather than single reference models, and creating curated benchmarks that avoid contamination from LLM training data.

Research on LLM-based ontology population addresses the A-box instantiation challenge through diverse approaches. Ciatto et al. [36] propose using LLMs as *oracles* for instantiating ontologies with domain-specific knowledge, querying models iteratively with templates derived from T-box structures to populate classes and properties. Experiments in the nutritional domain demon-

strate that structured prompts encoding ontological constraints guide population while reducing hallucinations, though quality metrics remain substantially lower than manual curation (5× improvement over baselines but requiring extensive post-validation). Sahbi et al. [169] compare semantic approaches with LLM-based extraction for populating ontologies from French classified advertisements. Their empirical comparison reveals complementary strengths: traditional semantic methods achieve higher precision through explicit knowledge-based analysis, while LLMs demonstrate greater flexibility handling diverse linguistic patterns and implicit relationships, albeit with lower precision and higher hallucination rates.

The hallucination challenge, identified as a critical limitation for high-precision applications like CH knowledge extraction, has motivated research on structured prompting strategies. Shyama et al. [183] develop pattern-based prompting for extracting A-box axioms, explicitly conveying ontology design patterns (hierarchical/taxonomy, n-ary relation, binary relation, attribute-value) in prompts to constrain LLM outputs. Their experiments demonstrate that pattern-based prompts achieve 100% recall for relationship assertions compared to 4.40% with conventional few-shot strategies, with F_1-scores ranging from 62.50% (class assertions) to 91.03% (property assertions). These results suggest that encoding ontological structure in prompts addresses LLM hallucination concerns, though performance remains variable across assertion types.

The question of human oversight in LLM-based ontology engineering directly bears on CH applications' requirements for semantic precision and provenance. Doumanas et al. [53] investigate collaborative workflows spanning a spectrum from human-centered to fully automated LLM-based ontology engineering. Their experiments on Search and Rescue domain ontologies reveal that while complete automation significantly accelerates development, human-LLM collaboration produces KGs with substantially higher precision (0.67-0.71) and recall (0.75-0.80) compared to fully automated approaches (preci-

sion 0.54, recall 0.63) when evaluated against expert-created reference KGs. The authors conclude that LLMs function most effectively as assistants within expert-driven workflows rather than autonomous ontology engineers, with optimal results achieved when domain experts validate LLM-generated components incrementally.

These findings carry implications for potential LLM integration in CH text-to-KG pipelines. The pattern of higher recall but lower precision observed across multiple studies ([36], [169], [183]) suggests that LLMs may address the scarcity of data endemic to specialized domains by generating candidate extractions requiring subsequent validation, potentially reducing manual annotation effort while maintaining quality through expert oversight. These observations suggest that LLMs might function as components within modular pipelines (similar to how projects combine rule-based extraction, NER tools, and custom validators) rather than as complete replacement systems, contributing extraction capabilities while remaining subject to the same validation requirements applied to other automated methods.

2.6 Defining Text-to-KG for Cultural Heritage

The preceding analysis enables the synthesis of a tentative formal framework characterizing text-to-KG approaches in CH contexts, which we will use from now on within this work. The term "text-to-KG" appears across CH literature with varying interpretations. Analysis of the eleven surveyed projects reveals that successful implementations consistently adhere to ontology-driven extraction: systems operate within predefined schemas rather than discovering structure from text. This observation aligns with the Closed KE paradigm and OBIE approaches characterized by Wimalasuriya et al. [215].

Based on the surveyed implementations, we define text-to-KG for CH applications as follows: text-to-KG refers to ontology-driven systems where ontologies serve dual purposes: **guiding the design** and implementation of extraction mechanisms, and **providing the formal structures** for represent-

ing extracted knowledge. This definition excludes systems that construct or extend ontologies during extraction; we assume target ontologies exist prior to extraction and remain static throughout the process. The task is populating these ontologies with **instances** and property **values**, and connecting those through available relationships within the ontology.[39]

Mathematical Formulation

Let T represent the set of all units of knowledge expressible from the source text. A unit of knowledge corresponds to one of the following structures:

- **Class membership assertion**: An entity e instantiates class C, expressible as $e \in C$ or in RDF as $(e, \texttt{rdf:type}, C)$
- **Binary relationship assertion**: Two entities e_1 and e_2 are related by property p, expressible as (e_1, p, e_2) where e_1, e_2 are entity URIs
- **Data property assertion**: An entity e is associated with a literal value v through property p, expressible as (e, p, v) where v is a literal (string, date, number, etc.)
- **Complex event structure**: In event-centric ontologies like CIDOC CRM, a unit of knowledge may correspond to a reified event structure connecting multiple entities through an intermediate event node

Let O represent the set of all valid knowledge configurations according to an input ontology, expressed in a formal knowledge representation language such as OWL or RDF Schema. Let K represent the KG produced by the extraction system.

The text-to-KG task can be formalized as a partial function:

$$f : T \rightharpoonup K \text{ where } K \subseteq O$$

[39] Throughout this work, text-to-KG will be used to refer to this definition. Alternative formulations such as "closed-ontology-based-text-to-KG-ontology-population" would have been more precise but less wieldy.

This formulation employs a partial function (denoted $\rightharpoonup$) rather than a total function to capture the observation that not all knowledge units in source texts are extracted by surveyed systems. Some textual information falls outside the scope of target ontologies, while other information may be present but unextractable due to linguistic complexity, ambiguity, or methodological limitations. The constraint $K \subseteq O$ ensures that all extracted knowledge conforms to the ontological model. We can also define four properties that characterize successful text-to-KG implementations:

- **Ontology Compliance**: All elements in K must satisfy the structural and semantic constraints defined in O. For example, if ontology O specifies that property p has domain class C_1 and range class C_2, then any triple $(s, p, o) \in K$ must have s instantiating C_1 and o instantiating C_2. This property ensures that extracted graphs remain semantically valid and can be integrated with other RDF data adhering to the same ontological framework.

- **Groundedness**: For every assertion $k \in K$, there must exist at least one knowledge unit $t \in T$ that justifies the inclusion of k in the output graph. Formally:

$$\forall k \in K, \exists t \in T : \text{grounds}(t, k)$$

  where $\text{grounds}(t, k)$ indicates that the textual knowledge unit t provides evidential support for the assertion k.

- **Appropriate Scope**: The partial function f should map knowledge units from T to K only when the corresponding knowledge is expressible within the conceptual scope of the ontology O. Knowledge that falls outside the domain coverage of the ontology should not be forced into ill-fitting ontological structures. Formally, if $t \in T$ is mapped to some $k \in K$, then the semantic content of t must be representable using the classes and properties defined in O.

- **Extraction Completeness**: Although not always achievable in practice, an aspirational text-to-KG system should extract all knowledge from T that is both (a) in the domain of O and (b) identifiable by the available extraction methods. This can be stated as:

$$\forall t \in T : (\text{expressible}(t, O) \wedge \text{identifiable}(t)) \Rightarrow \exists k \in K : \text{grounds}(t, k)$$

2.6.1 The Dominant Pipeline Architecture

Seven of the eleven surveyed projects employ variations of a modular pipeline architecture, decomposing the extraction task into sequential sub-tasks. This section synthesizes the common architectural pattern identified across these implementations.

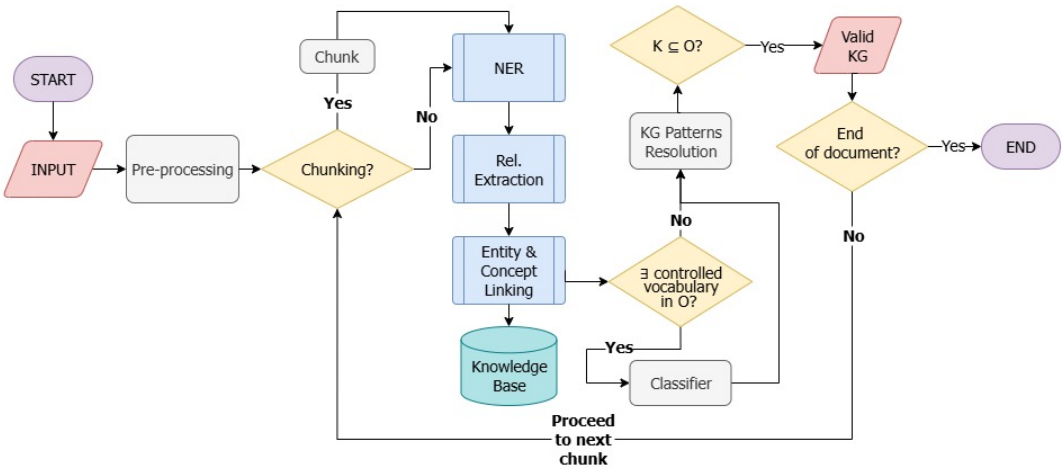


Figure 2.1: Flowchart of the modular text-to-KG pipeline architecture observed across surveyed CH projects

Figure 2.1 presents a flowchart summarizing the components that typically constitute this pipeline based on analysis of BiographySampo, WarSampo, MMKG, MusicBO, Bernoulli-Euler, Viewsari, InTaVia and existing literature reviews [132]. While individual implementations vary in component ordering, implementation technologies, and the presence or absence of particular stages, this architecture represents the dominant pattern observed across CH text-to-KG systems. The typical pipeline observed in surveyed projects includes the

following stages:

1. **Text Preprocessing**: Sentence segmentation, tokenization, and potentially coreference resolution to identify entity mentions that refer to the same real-world entity across the document. Preprocessing may also include document structure analysis for sources like scholarly articles where section headers, footnotes, and bibliographic references require specialized handling. Projects processing historical texts implement additional steps such as medieval abbreviation expansion, orthographic normalization, or handling of multilingual text where Latin is interspersed with vernacular languages.

2. **Named Entity Recognition (NER)**: Identification and classification of entity mentions according to predefined types. General-purpose NER systems typically recognize Person, Location, Organization, and occasionally Date or Quantity, but CH applications frequently extend these taxonomies with domain-specific types such as musical works, sensory descriptors, occupation categories, or professional relationships. Multilingual settings introduce additional complexity, as entity recognition must operate across languages with different morphological properties.

3. **Entity Linking**: Disambiguation of entity mentions by mapping them to canonical identifiers in knowledge bases or authority files. This process resolves cases where surface forms are ambiguous or where different surface forms refer to the same entity. Entity Linking emerged as the most frequently addressed bottleneck across surveyed projects (Section 2.4), particularly for historical texts containing long-tail entities absent from large-scale knowledge bases, which require integration with specialized authority files or manual curation workflows.

4. **Relation Extraction**: Identification of semantic relationships between entities that correspond to properties defined in the target ontology. Surveyed projects employ diverse approaches ranging from pattern-based

template matching to neural relation classifiers and frame-semantic annotation.

5. **Event Detection and Reification**: Projects employing event-centric ontologies such as CIDOC CRM require a dedicated step that aggregates multiple entities and relations into complex event structures—for instance, production events linking creator, created object, time, and place. This component also incorporates temporal and spatial extraction to recognize and normalize date expressions and geographic references, a capability observed as essential across multiple CH applications.

6. **Graph Assembly and Validation**: Extracted entities, relations, and events are assembled into coherent graph structures. Validation checks ensure ontology compliance, verifying that property domains and ranges are satisfied, required properties are present, and cardinality constraints are respected. Approaches range from formal constraint languages such as SHACL to manual expert review.

7. **Provenance Annotation**: Links between extracted assertions and source text spans are established, recording which sentences or passages provided evidence for each triple. Systems may also record extraction confidence scores, alternative interpretations considered but rejected, and metadata about the extraction process itself. Several projects implement comprehensive provenance tracking using PROV-O or named graphs to distinguish assertions derived from different sources.

The modular architecture enables systems to leverage specialized techniques for each subtask—pattern matching for preprocessing, statistical models for NER, graph-based algorithms for Entity Linking, neural classifiers for Relation Extraction—while maintaining clear interfaces between components that facilitate debugging, evaluation, and incremental improvement. This architectural pattern's prevalence suggests its suitability for CH applications despite the domain-specific adaptations required for each component.

Beyond architectural similarities, the surveyed projects converge on a shared extraction paradigm: all eleven implementations adopt Closed Knowledge Extraction, operating within predefined ontological frameworks rather than discovering structure from text. This uniformity contrasts with general-domain text-to-KG research, where Open Information Extraction maintains a significant presence. Several factors account for this preference in CH contexts: institutional integration requirements demand compatibility with established metadata schemas and controlled vocabularies; ontological constraints provide validation mechanisms such as SHACL shape checking that support the semantic precision expected in scholarly applications; and explicit schema definitions facilitate provenance tracking between extracted triples and source passages, aligning with CH practitioners' emphasis on traceability.

2.7 Conclusion

This chapter examined text-to-KG extraction in the CH domain through integrated analysis of implemented projects, domain-specific challenges, and emerging methodological trends. Section 2.6 synthesized these empirical observations into a formal framework defining text-to-KG for CH applications, characterizing the dominant modular pipeline architecture, and explaining the paradigmatic preference for Closed KE approaches. Together, these analyses directly address the three research questions established in Section 2.1:

- **RQ1.1: Which projects have successfully generated KGs from CH texts using text-to-KG methodologies?** The analysis identified eleven projects spanning 2015-2025 that meet the established inclusion criteria. These projects span diverse CH subdomains (biographical heritage, parliamentary records, military history, olfactory heritage, musical encounters, art history, scientific correspondence, legal documents) and demonstrate varying technical maturity. Early implementations such as WarSampo (2014-present) relied primarily on pattern-based extraction with domain-specific ontologies, while recent projects such as Odeuropa

(2020-2023) and InTaVia (2021-2024) employ transformer-based architectures with multilingual capabilities—either with custom training or by relying on ready-to-use tools. The comparative analysis revealed that seven of eleven projects employ variations of the modular pipeline architecture synthesized in Section 2.6.1, with modifications addressing specific domain requirements rather than fundamental architectural redesign.

- **RQ1.2: What challenges emerge when applying KE techniques to CH documents?** The challenge analysis identified bottlenecks operating at two levels: document characteristics and pipeline component performance. Document-level challenges include OCR quality degradation, historical language variation, multilingual corpora, and semantic change across temporal contexts. Component-level bottlenecks center on entity linking, which emerged as the most frequently addressed challenge across projects, followed by NER and relation extraction. The prevalence of long-tail entities in CH corpora limits the effectiveness of general-purpose entity linking tools trained primarily on Wikipedia and news corpora. Data alignment challenges manifest across three dimensions: T-Box alignment (schema extension for novel entity types), A-Box alignment (deduplication and consistency checking), and URI alignment (entity resolution). Coreference resolution constitutes a bottleneck only when KG scope requires entity resolution across discourse boundaries; projects with sentence-level or event-level extraction scope avoid this challenge through design choices.

- **RQ1.3: What methodological trends address the challenges identified in CH text-to-KG implementations?** Literature analysis of 227 papers published 2020-2025 reveals concentration in three technological approaches: LLMs (44 papers mentioning in titles), KG embeddings (11 papers), and Graph Neural Networks (9 papers). The publication acceleration from 2023-2025 aligns with the timeline of surveyed

projects. The literature addresses specific bottlenecks through diverse approaches: generalist NER tools for domain adaptation without extensive retraining; automatic dataset generation techniques for specialized entity types; temporal-aware historical NER approaches for language variation; generative approaches to relation extraction; LLM-based unsupervised relation extraction with larger context windows; approaches for unseen entity linking; and LLM-based evaluation frameworks as alternatives to traditional metrics.

Beyond answering these specific questions, the synthesized framework (Section 2.6) establishes evaluation criteria for text-to-KG systems through four essential properties: ontology compliance ensures extracted graphs satisfy ontological constraints; groundedness prevents hallucination by requiring textual evidence for all assertions; appropriate scope respects ontological boundaries; and extraction completeness serves as an aspirational goal. These properties, derived from analysis of successful implementations, provide the foundation for evaluating both existing systems and the novel methodologies developed in subsequent chapters.

Three methodological patterns emerge from the comparative analysis that characterize successful CH text-to-KG implementations. First, **hybrid architectures combine multiple approaches rather than relying on single methodologies** (e.g., MMKG). MusicBO integrates neural AMR parsing with back-translation quality control. This suggests that architectural modularity provides flexibility that single-approach optimization cannot achieve. Second, domain-integrated design requires **integration of domain expertise** throughout the pipeline rather than as post-processing validation. Odeuropa's frame-semantic annotation schemes, developed in parallel with ontology engineering, enable direct mapping from extracted frame elements to ontological concepts. Third, **provenance maintenance** distinguishes CH applications from general-purpose extraction systems. The progression from BiographySampo's pattern-based extraction to InTaVia's dual-framework NLP

pipeline illustrates technical maturation through accumulative sophistication rather than linear replacement. Recent projects maintain rule-based components for domain-specific tasks while leveraging neural architectures for general linguistic processing. The emergence of generalist tools such as GLiNER, which recognizes arbitrary entity types through natural language descriptions, offers promise for CH applications where entity taxonomies are diverse and domain-specific. Similarly, zero-shot entity linking approaches address the challenge of linking mentions to entities not seen during training. However, several challenges remain inadequately addressed. The long-tail entity problem, while better quantified through empirical studies such as KE-MHISTO (demonstrating 30% NIL entities in historical music periodicals versus near-zero in contemporary benchmarks), continues to limit effectiveness of general-purpose tools. Gender bias in both historical texts and knowledge bases creates systematic representation gaps (female historical figures constituting 50-55% NIL entities versus 27-28% for males in music corpora) that technical solutions alone cannot resolve. Data scarcity remains a constraint, though approaches such as automatic dataset generation and few-shot learning show promise. The heterogeneity of evaluation reflects the absence of standard KG generation protocols, with projects employing various validation approaches adapted to specific motivations and audiences.

The analysis suggests considerations for CH practitioners implementing text-to-KG approaches. Architectural modularity enables adaptation to specific corpus characteristics and research requirements, as demonstrated by successful projects combining general-purpose NLP tools with domain-specific components. Scope-appropriate design aligns KG extraction boundaries with research questions. Provenance maintenance requires explicit links between extracted knowledge and source materials by design rather than as afterthought. Evaluation planning must address scale constraints: MusicBO's back-translation approach addresses the impossibility of manually evaluating tens of thousands of KGs, though closed KE cases extracting few triples per

text may require alternative approaches. Human evaluation remains the gold standard, although resource constraints require complementary automated validation.

This survey faces limitations reflecting both methodological constraints and structural characteristics of CH research contexts. The project analysis centers around Western, particularly European, initiatives, likely due to funding accessibility and documentation practices. Most identified projects benefited from European H2020 funding or equivalent institutional support. The text-to-KG pipeline requires substantial prerequisites: digitized collections, technical infrastructure, and sustained funding for interdisciplinary teams, creating systematic barriers for resource-constrained institutions. The challenge analysis is based primarily on documented project experiences, potentially missing domain-specific obstacles that prevented projects from reaching completion or publication.

The convergence of improved multilingual language models, better understanding of domain-specific challenges, and growing availability of CH datasets suggests potential for advances. However, progress requires continued collaboration between technical researchers and domain experts. Future research should prioritize documenting projects from diverse cultural contexts, developing lighter approaches that can operate with limited training data, creating evaluation frameworks that balance technical rigor with domain-specific requirements, and addressing systematic biases in source materials and knowledge bases. The measure of success for text-to-KG systems in CH contexts is not technical metrics alone, but their capacity to enable new forms of scholarly inquiry while preserving the richness of the input source and, of course, the provenance.

Chapter 3

Towards LLM-assisted pipelines for text-to-KG

"It is not right to say 'no' before you have tried."

— Roberto Busa (1980)

The survey in the previous chapter highlighted that Large Language Models (LLMs) are becoming a potential tool for Knowledge Extraction. Our investigation addresses three interrelated problems that emerge from integrating LLMs into ontology-driven KE pipelines. First, the generation of Knowledge Graphs (KGs) from text involves numerous standards and variables that change across domains, raising questions about whether optimization should target end-to-end LLM approaches or pipeline modules. While LLMs demonstrate few-shot capabilities that suggest promise for domain adaptation, current limitations in ontology alignment and RDF syntax generation require careful consideration of where LLM integration provides advantages [164]. Second, depending on target ontology requirements, handling implicitness in source texts becomes necessary; LLMs theoretically possess inferential capabilities that could address this requirement, though the extent and reliability of this capability requires empirical investigation [186]. Third, the generative nature of LLMs for compound tasks like text-to-KG makes character-level evaluation impractical, necessitating evaluation frameworks that account for semantic correctness beyond surface-level string matching [91].

Testing these problems comprehensively requires extensive experimental validation. While the following chapters present empirical investigations, we acknowledge that current research has not completely established the boundaries of LLM capabilities in this domain. Given this uncertainty, we rely on experimental literature on related topics while recognizing potential biases in selecting supporting evidence. We attempt objectivity and avoid claims that generalize beyond the scope of the proofs of concept elaborated in subsequent chapters.

This chapter contains three sections. In the first section, we explore literature to understand how LLMs fit within text-to-KG pipeline modules. In the second section, we present an analysis of LLM capabilities in handling implicit and explicit information, a feature that demonstrates potential advantages for KE compared to non-generative models in Cultural Heritage (CH) domain (and subdomains). The third section defines the scope and boundaries for LLM integration within text-to-KG pipelines, identifying key problems that motivate the experimental work in subsequent chapters.

3.1 LLM-assisted text-to-KG modules

The application of LLMs to **Named Entity Linking** (NER) has yielded mixed results across different evaluation settings. In evaluations on benchmark datasets including CoNLL2003 [198] and OntoNotes5.0 [214], LLMs show performance below baseline for the NER task against bidirectional encoders such as BERT [211, 225]. One of the main reasons is a tendency to identify entities in null inputs (i.e., hallucinations or false positives) [211]. The output format of LLM-based NER also presents challenges. NER is traditionally a sequence labeling task; an LLM can either rewrite the entire input with annotations (at the cost of potentially providing distorted text) or provide a list of outputs (with no index information) as a list, JSON, or other format. Wang et al. [211] address the format issue by transforming NER into a generation task where special tokens mark entity boundaries (e.g., transforming

"Columbus is a city" into "Columbus## is a city" to identify location entities), which reduces the generation difficulty as the model only needs to mark entity positions while copying remaining tokens. To mitigate hallucination, they propose a self-verification strategy that prompts the LLM to validate whether extracted entities belong to the specified entity tags, acting as a regulatory mechanism against over-confident predictions. The advantage of LLMs lies in low-resource scenarios where LLMs show better performance than state-of-the-art sequence-labeling NER models [211, 160]. An LLM output can also be corrected automatically by asking the model to cross-check the result of its first output against the input text or to correct or enrich the results of a sequence-labeling NER model to check for potential false negatives [211].

On **Relation Extraction**, BERT-based models dominate benchmarks. Their effectiveness stems from pre-training objectives: Masked Language Modeling and Next Sentence Prediction align closely with identifying relationships between entities as well as identifying the entities themselves [48]. LLMs contribute approximately 25% to state-of-the-art RE results, with notable performance in specific scenarios rather than across-the-board superiority [48]. In specialized domains where annotated training data is scarce, LLMs demonstrate superior performance compared to traditional models, as they can capture nuanced semantic connections that discriminative models miss or relations that, compared to BERT models, lie in the long tail of training data. Their ability to process longer contexts makes them suitable for document-level RE, where relationships span entire documents rather than single sentences [48].

LLMs have been characterized as implicit knowledge bases capable of storing and retrieving factual information through natural language queries, suggesting potential application for **Entity Linking** [152]. However, their reliability as entity linkers depends on entity frequency in pre-training data. Performance degrades substantially for entities in the long tail of the popularity distribution, particularly in specialized domains distant from web-based training corpora [210]. Evaluation on KE-MHISTO, a multilingual benchmark

for entity linking in historical musical heritage documents, demonstrates this limitation: state-of-the-art LLMs achieve F_1 scores between 0.48 and 0.61 on English historical periodicals and 0.33 to 0.51 on Italian materials, with performance declining further when stratified by entity popularity [81]. Analysis reveals that LLMs correctly link popular entities while failing systematically on rare entities, with performance approaching near-zero F_1 scores for entities below the 50th popularity percentile [81].

Despite limitations in direct entity linking, LLMs demonstrate utility as re-ranking components within hybrid pipelines. When provided with candidate entities from knowledge base retrieval systems, LLMs can leverage source document context to disambiguate between alternatives [210]. This addresses the tendency of frequency-based methods to select popular entities regardless of context. Experiments with NIL[1] entities prediction reveal trade-offs between precision and recall across different LLMs, with some models achieving high precision (0.82) but low recall (0.35) on NIL entities, while others exhibit inverse patterns [81].

LLMs can be adapted as **Classifiers** by leveraging class features and examples through few-shot learning, enabling task adaptation without extensive fine-tuning [27]. This capability allows LLMs to perform classification tasks across diverse domains with minimal labeled examples, circumventing the need for task-specific training data. In-context learning enables LLMs to generalize from a small number of demonstrations provided in the prompt, making them suitable for classification problems in low-resource settings or specialized domains where annotated datasets are unavailable. In text-to-KG, this proves useful in vocabulary alignment tasks, such as mapping terms to classes of CIDOC CRM [218].

Beyond direct classification, LLMs can augment existing classifiers by providing explainability for predictions, enhancing interpretability in Knowledge Extration (KE) pipelines [134]. Retrieval-augmented generation (RAG) ap-

[1]NIL is usually interpreted as "Not In Lexicon" in the Entity Linking task. Its original meaning is a synonymous of *nothing* as reported by the Oxford Dictionary

proaches further enhance classification performance by grounding LLM predictions in retrieved contexts [37]. However, the application of LLMs as few-shot classifiers requires consideration of inherent biases that stem from training data dominated by English-language and Western sources [124, 192].

3.2 LLMs and Implicit Information

Real-world texts convey information both explicitly and implicitly, with the latter requiring inferential processing and contextual reasoning to derive intended meaning [186].[2] Implicit information arises when facts are conveyed indirectly through linguistic mechanisms rather than direct statements. For instance, the sentence "Zuhdi attends church every Sunday" implies religious affiliation through contextual inference rather than explicit declaration [186]. This distinction presents challenges for knowledge extraction systems, as traditional methods predominantly rely on explicit textual markers to identify entities and relationships.

The degree of implicitness directly impacts extraction difficulty and model certainty. Consider three statements conveying identical information with varying implicitness:

1. "Gaia works as a doctor at City Hospital" (explicit)
2. "Gaia wears a white coat and sees patients daily" (moderately implicit)
3. "Gaia ran through the emergency room corridor, quickly reviewing charts" (highly implicit)

While the first statement maps directly to a subject-predicate-object structure, the latter two require progressively more inference, introducing uncertainty that compounds as explicitness decreases [186]. The third statement demonstrates aleatoric uncertainty—inherent linguistic ambiguity—as multiple roles could plausibly explain the described behavior without additional context.

[2]This section draws from prior work by the author: Stramiglio, A., Schimmenti, A., Pasqual, V., van Erp, M., Sovrano, F., & Vitali, F. *Explicit vs. Implicit Biographies: Evaluating and Adapting LLM Information Extraction on Wikidata-Derived Texts* (2025, arXiv:2509.14943 [cs.CL].) [186]

LLMs demonstrate systematic difficulties in extracting information from implicit contexts. We perform a test on three LLMs (Phi-1.5,[3] LLaMA 3.2 1B,[4] DeepSeek-R1-Distill 1.5B[5]) over a dataset of implicit and explicit sentences describing a person's occupation. Evaluation on synthetic biographical datasets reveals that models exhibit significantly higher semantic distance between predictions and ground truth for implicit versus explicit verbalizations (Wilcoxon signed-rank test, $p < 0.05$) [186]. Models produce substantially higher failure rates when processing implicit texts (14.6% non-responses) compared to explicit texts (1.3%), indicating fundamental limitations in handling indirect information [186]. Performance degradation is particularly pronounced for relation extraction, where implicit statements require commonsense reasoning and domain knowledge beyond surface-level linguistic patterns. Models frequently retrieve hypernyms rather than specific hyponyms (e.g., "actor" instead of "television actor"), demonstrating reduced precision on implicit extractions.

Table 3.1 presents results from these three LLMs in the 1-1.5B parameter range evaluated on extraction tasks using texts derived from Wikidata biographies. Pre-trained models without specialized training achieve accuracy above 88% on explicit information extraction but only 58-71% on implicit information extraction, representing performance drops of 17-31 percentage points [186].

However, fine-tuning on implicit data patterns substantially improves model performance on implicit reasoning tasks. Models trained on both explicit and implicit data achieve accuracy exceeding 90% on implicit extraction tasks, compared to 58-71% for models trained exclusively on explicit data [186]. This improvement is consistent across model architectures within the 1-1.5B parameter range, suggesting that the difficulty stems primarily from insufficient exposure during pre-training rather than architectural limitations—epistemic rather than aleatoric uncertainty.

[3] https://huggingface.co/microsoft/phi-1_5
[4] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct
[5] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

Table 3.1: Performance comparison of LLMs on implicit versus explicit information extraction tasks. Results demonstrate model size effects on handling implicit reasoning.

Model	**Size**	**Explicit**	**Implicit**
Phi-1.5	1.5B	0.889	0.581
LLaMA 3.2	1B	0.888	0.716
DeepSeek-R1-Distill	1.5B	0.883	0.671
After fine-tuning on implicit data:			
Phi-1.5 (FT)	1.5B	0.896	0.925
LLaMA 3.2 (FT)	1B	0.892	0.933
DeepSeek-R1-Distill (FT)	1.5B	0.900	0.907

These findings have implications for text-to-KG pipelines in CH domains. While small models (1-8B parameters) demonstrate substantial performance gaps on implicit reasoning without specialized training, larger models may exhibit stronger capabilities for handling indirect information through broader exposure to diverse linguistic patterns during pre-training. For CH domains where information is frequently conveyed implicitly, such as historical documents or domain-specific corpora, targeted fine-tuning remains necessary for smaller models. The magnitude of required adaptation likely decreases with model scale, as larger models possess more robust representations for inferential reasoning [186].

The ability to handle implicit information represents a potential advantage of LLMs over traditional NER and RE systems in CH contexts. Historical texts often convey relationships and entity properties through narrative context rather than explicit statements. For instance, family relationships might be inferred from patronymic naming conventions, professional roles from described activities, or temporal information from contextual markers. Traditional pipeline components trained primarily on explicit relationship patterns may fail to extract such information, whereas LLMs trained on diverse textual corpora may possess the inferential capabilities required for such extraction—provided they receive appropriate task-specific adaptation or operate at sufficient scale.

3.3 LLMs within text-to-KG: Defining scope and boundaries

The analysis of LLM capabilities across text-to-KG modules reveals a pattern of specialized rather than universal applicability. LLMs do not uniformly outperform traditional approaches across all pipeline components; instead, they demonstrate advantages in specific contexts that correspond to characteristics of the CH domain identified in Chapter 2.

3.3.1 Advantages

Three scenarios emerge where LLMs provide advantages over traditional pipeline components. First, in low-resource domains where annotated training data is scarce or unavailable, LLMs leverage few-shot learning to perform extraction tasks without extensive domain-specific training [27]. This aligns with CH challenges where entity types and relationship schemas vary across projects and corpora, making creation of large annotated datasets impractical. Second, when extraction requires long-range contextual reasoning beyond sentence boundaries, LLMs' capacity to process extended contexts enables document-level extraction where traditional sentence-level models fail. This addresses CH documents where information is distributed across paragraphs or sections rather than concentrated in individual sentences. Third, in handling implicit information where facts must be inferred from narrative context rather than explicit statements, LLMs demonstrate capabilities that traditional pattern-based or sequence-labeling approaches lack, provided appropriate scale or fine-tuning.

3.3.2 Limitations

Three categories of limitations constrain LLM application in text-to-KG pipelines. First, performance limitations manifest in tasks where discriminative models trained on large annotated datasets outperform generative approaches. Entity linking for long-tail entities, where LLMs rely on memorization of training data rather than reasoning about entity properties, exempli-

fies this limitation. The KE-MHISTO results demonstrate systematic failures on entities below median popularity, suggesting that LLMs' knowledge base characteristics poorly match the distribution of entities in CH corpora [81]. Second, output format constraints arise from the inherent tension between generative flexibility and structured output requirements. While LLMs can generate natural language descriptions and have been extensively trained for JSON structures (especially for function and tool call in latest iterations) , ensuring syntactic correctness of RDF serializations or maintaining alignment with target ontologies requires additional verification layers [164]. Third, evaluation challenges emerge from the disconnect between generative output and traditional sequence-labeling evaluation frameworks, particularly when LLMs enrich or transform entity mentions during extraction.

3.3.3 Hybrid architectures as pragmatic solutions

The limitations and advantages suggest that hybrid architectures combining LLMs with traditional components offer pragmatic solutions for CH text-to-KG pipelines. Rather than replacing entire pipelines with end-to-end LLM approaches, integration strategies position LLMs where their capabilities address specific bottlenecks while retaining traditional components for tasks where they maintain advantages. This aligns with patterns observed in recent CH projects surveyed in Chapter 2, where successful implementations combine multiple methodological approaches rather than relying on single techniques [137].

Potential integration patterns include: LLMs as enrichment components that augment outputs from traditional NER or RE systems, adding inferred information or resolving ambiguities through contextual reasoning; LLMs as verification components that validate or correct outputs from traditional pipelines, reducing false positives or false negatives; LLMs as re-ranking components that disambiguate between candidate entities or relationships using extended context; and LLMs as low-resource alternatives deployed specifically for entity types or relationships where annotated training data is unavailable.

3.3.4 Open problems for LLM integration

Three problems emerge from this analysis that motivate the experimental investigations in subsequent chapters. These problems address **RQ2**: *How can Large Language Models be integrated into ontology-driven Knowledge Extraction pipelines for Cultural Heritage texts, and what are the limitations, requirements, and trade-offs of such integration?*

Problem 1: Context-aware task formulation. The recurring architecture of ontology-driven text-to-KG tasks need not be overhauled by LLMs, but LLMs change how tasks can be formulated. Traditional pipeline components perform isolated extraction tasks—NER identifies entity boundaries and types, RE classifies relationships between entity pairs—without access to broader context about entity roles or document structure. LLMs potentially enable context-aware formulations where extraction tasks incorporate additional constraints or requirements derived from target ontology structure or domain knowledge. For instance, NER labels might be conditioned not only on entity type definitions but also on expected roles within target relationship schemas. Consider the example:

> [Andrea Costa] was born in Imola on November 29, 1851, to Pietro and Rosa Tozzi in a practicing Catholic family of modest conditions.[6]

An LLM can be instructed to perform NER while also formatting names in their fullness (e.g., given name, surname, and additional names). Within this instruction, identifying Andrea Costa's father would require inferring that "Pietro" shares the surname "Costa" based on familial relationship and cultural naming conventions, producing "Pietro Costa" as the entity mention despite "Costa" appearing nowhere near "Pietro" in the source text. This task implies both handling implicit relationships and applying cultural knowledge that LLMs may possess through pre-training. Whether such context-aware for-

[6]The quote is a translation by the author from Incipit of Andrea Costa's biography from the Italian National Archiving System (SAN)

mulations provide practical advantages over traditional pipeline architectures combining separate NER, coreference resolution, and relationship extraction components requires empirical investigation.

Problem 2: Evaluation frameworks for generative extraction. Given the generative nature of LLMs, evaluating outputs within sequence-labeling frameworks proves impractical. Traditional NER evaluation relies on exact span matching between predicted and gold-standard entity boundaries. In the Pietro Costa example, an LLM correctly inferring "Pietro Costa" as an entity mention would fail string-matching evaluation against a gold standard annotating only "Pietro" at the token level. The LLM is not being evaluated on its ability to generate an entity entry following specified rules, but rather on its ability to recall exact tokens from which it inferred the full entity name. This suggests that traditional precision and recall metrics calculated through string matching may systematically underestimate LLM performance on extraction tasks that involve inference or normalization. However, alternative evaluation approaches must balance flexibility in matching generated outputs against gold standards with rigor in assessing correctness. Qualitative assessment can identify correct extractions that automated metrics miss, but cannot scale to large-scale evaluation required for systematic comparison across approaches or models.

Problem 3: Ontology alignment and knowledge graph construction. While LLMs demonstrate capabilities in extraction tasks producing entity mentions and relationship tuples, translation of extracted information into KG structures aligned with target ontologies presents additional challenges. Ontology-driven KE requires not only identifying entities and relationships in text but also mapping them to ontology classes, properties, and constraints. This includes URI assignment for entity resolution, property selection from target vocabularies, and enforcement of domain and range constraints. Current LLMs show limitations in generating syntactically correct RDF serializations [218] and maintaining consistency with ontological constraints without

extensive prompting or verification. Whether LLMs can be effectively integrated into the ontology alignment and KG construction phases of text-to-KG pipelines, or whether these phases require traditional symbolic approaches, remains an open question requiring investigation.

3.4 Conclusions

The problems identified in this section will be seen throughout the experimental investigations throughout Chapter 4 to Chapter 8.

These problems cannot be fully resolved within the scope of a single dissertation. Rather, the experimental work provides empirical evidence regarding trade-offs and limitations of LLM integration in specific CH contexts, contributing to broader understanding of where and how LLMs fit within text-to-KG pipelines. The findings aim to inform practitioners implementing similar approaches in CH domains while identifying directions for future research addressing unresolved challenges.

Chapter 4

Knowledge Extraction on Archival Finding Aids

Chapter 2 and Chapter 3 established the boundaries of how Large Language Models (LLMs) can be implemented within the text-to-KG (Knowledge Graph) task. In this chapter, we present a proof of concept implementing an ontology-driven, LLM-based text-to-KG pipeline on archival finding aids. The pipeline translates biographical notes within archival descriptions into RDF triples aligned with the Records in Contexts Ontology (RiC-O), addressing the challenge of making latent contextual information computationally accessible within archival collections.[1]

The remainder of this chapter is organized as follows: Section 4.1 presents the archival context and identifies the computational challenge posed by narrative descriptions in finding aids. Section 4.2 describes the Records in Contexts Ontology as the target knowledge representation framework. Section 4.3 details the methodology, including the iterative workflow and evaluation framework. Section 4.4 presents the pipeline implementation and case study. Section 4.6 discusses findings and future developments. Section 4.7 concludes the

[1]This chapter draws from prior work by the author: Giagnolini, L., Schimmenti, A., Bonora, P., & Tomasi, F. (2025). *Expliciting Contexts: Semantic Knowledge Extraction from Traditional Archival Descriptions. Umanistica Digitale*, 9(20), 115–144. The author mainly contributed to the design and implementation of the knowledge extraction process. Dr. L. Giagnolini, digital humanist and archivist, was responsible for the conceptualization of the work and the methodology, and operated as the domain expert [78]

chapter.

4.1 Background

Archives constitute repositories of documentary evidence recording human activities across time, preserving the material traces of administrative, legal, social, and cultural processes. According to the International Council on Archives (ICA), an archive contains "the whole of the records, regardless of form or medium, organically created andor accumulated and used by a particular person, family, or corporate body in the course of that creator's activities and functions".[2] The principle of provenance is fundamental to archival science, as documents within an archive derive their meaning from their original context of creation and use, necessitating a description that preserves these contextual relationships [199].

Archival finding aids serve as the primary instruments for intellectual control and access to archival collections. These structured documents describe archival materials at multiple levels of granularity, from entire finds to individual items, providing researchers with essential information about content, context, and access conditions. The International Standard Archival Description (General), or ISAD(G), provides standardized elements for multilevel description including identity statements, context areas, content and structure descriptions, conditions of access and use, and notes fields.[3] Within these structured frameworks, archivists embed substantial contextual knowledge in textual form: biographical sketches of record creators, administrative histories of corporate bodies, custodial histories tracing document transmission, and scope notes describing intellectual content and historical significance [230].

These narrative components of finding aids represent a particular form of professional discourse, combining factual documentation with interpretative synthesis. The biographical note for a personal archive, for instance, weaves

[2] https://www.ica.org/resource/isadg-general-international-standard-archival-description-second-edition/

[3] https://www.ica.org/resource/isadg-general-international-standard-archival-description-second-edition/

together dates and places with assessments of historical significance, professional relationships, and cultural impact. Such texts encode not merely isolated facts but networks of relationships: between persons and institutions, between documents and the events they record, between archival materials and the broader historical contexts within which they acquire meaning. An archivist describing the papers of a political figure necessarily situates that individual within networks of political parties, governmental bodies, social movements, and historical events [202, 230].

However, these rich narrative descriptions remain largely opaque to computational processing in current digital archival systems. While the structured elements of ISAD(G)—dates, extent, reference codes—translate readily into database fields and XML schemas, the narrative components persist as unstructured text strings [230]. This limitation becomes particularly acute in Linked Open Data implementations, where archives publish structured metadata as RDF triples, while biographical notes and historical contexts remain literal values, unable to participate in the semantic reasoning and cross-collection linking that LOD enables [202].

Text-to-KG approaches offer a pathway to address this computational opacity. By systematically transforming narrative descriptions into semantic triples, archival institutions can extract both explicit and implicit relationships within textual content, enhancing the depth of contextual knowledge, enabling more sophisticated searches, and providing support for disambiguating reference entities [199, 77]. In this context, text-to-KG assigns relational semantics among records and finding aids, conveying specific informational components such as references to institutions, people, events, places, and temporal coordinates through individual triples.

This process constitutes, essentially, an interpretative act [41]: the embedding of contextual information derived from narrative text must be explicitly marked with provenance metadata distinguishing computationally derived knowledge from authoritative archival assertions. This can be achieved through

named graphs [33] or reification mechanisms that preserve the distinction between original descriptions and extracted information [202].

4.2 The Records in Contexts Ontology

The Records in Contexts-Ontology (RiC-O) constitutes an OWL ontology developed by the International Council on Archives Expert Group on Archival Description (ICA EGAD) for describing archival record resources.[4] As the formal representation of the Records in Contexts Conceptual Model (RiC-CM), RiC-O provides a structured vocabulary and formal rules for creating RDF datasets that describe archival materials in machine-readable format.

The ontology defines 107 classes organized hierarchically, including core entities such as `Record Resource`, `Agent`, `Activity`, and `Place`, along with supplementary classes for handling specific descriptive needs. The ontology implements RiC-CM relations through both binary properties and *n*-ary relation classes, enabling the representation of documented relationships with associated attributes such as dates, certainty, and provenance metadata.

In the context of biographical information, RiC-O provides specific classes and properties for representing the relationships between persons, corporate bodies, and their associated activities and records. The ontology defines `Agent` as a superclass with `Person` and `Corporate Body` as subclasses, enabling the representation of record creators and their contextual relationships.

RiC-O employs *n*-ary relation classes to model complex relationships with associated attributes. The `EventRelation` class serves as a general mechanism for describing any event involving entities. More specialized relation classes provide fine-grained representation of specific relationship types: `AgentToAgentRelation` models relationships between two agents; `PositionHoldingRelation` represents an entity holding a particular role or position; `PlaceRelation` connects places to other entities. Additional relation classes address specific aspects of biographical context, including

[4] `https://ica-egad.github.io/RiC-O/`

`FamilyRelation` for kinship connections and `TeachingRelation` for educational relationships. These *n*-ary classes enable the attachment of temporal, evidential, and descriptive attributes to relationships, supporting nuanced representation of biographical information. However, some relations are conceptualized as plain object properties, such as `rico:hasBirthPlace` and `rico:hasBirthDate` (between a `rico:Person` and a `rico:Date`).

4.3 Methodology

To demonstrate text-to-KG feasibility for archival finding aids, we adopted an iterative approach structured as follows:

1. **Ontology analysis**: Analysis of the RiC-O structure, patterns, and knowledge modeling approach to ensure each step of the text-to-KG pipeline aligns with the ontological framework.

2. **Document analysis**: Examination of document format, textual boundaries, length characteristics, and token distribution within the finding aids.

3. **Iterative testing**: Development of modular pipeline components tested individually to ensure outputs progressively match ontological requirements. Once individual patterns demonstrate satisfactory performance, the full pipeline is tested on the complete input.

4. **Evaluation**: Assessment of outputs through both quantitative metrics and qualitative analysis, employing domain expert evaluation given the focused scale of the examined finding aid.

4.3.1 Evaluation Framework

As discussed in Chapter 3, generative KE outputs present distinct evaluation challenges compared to traditional extraction approaches. The output labels and format may differ from expected forms while remaining semantically valid. Given the complexity of the KE task and the precision requirements for RDF

triple generation, we adopted a three-level evaluation framework examining structural, informational, and interpretative dimensions through both quantitative metrics and qualitative assessment. Given the focused scope of this proof of concept, manual assessment of model performance by a domain expert is feasible and necessary for validating semantic accuracy and interpretative fidelity.

Structural evaluation assesses how well outputs adhere to the intended schema design, examining both overall performance and specific event types. We focus on two criteria: schema adherence, checking if outputs follow the predefined event schema format and requirements; and consistency, verifying that similar information is represented uniformly across different event types and that structural patterns are maintained. This evaluation precedes the deeper analysis of KE quality and consists in counting the number of events produced corresponding to the schema.

Information extraction performance is evaluated using three metrics: Precision, Recall, and F_1 score. These metrics are calculated using a 2x2 confusion matrix that categorizes results into four outcomes:

- True Positive (TP): Information is correctly identified and accurately reported in the output matching its presence in the input. The information must be both present and classified correctly;

- True Negative (TN): Information is correctly identified as absent from both the input text and the output;

- False Positive (FP): Information is reported in the output but is either absent from or different in the input text with a wrongly executed inference. What the literature usually refers to as "hallucination" is defined as a FP in this context;

- False Negative (FN): Information is present in the input text but not reported in the output.

We report confusion matrices for overall results and per-class. Precision measures the proportion of correct predictions among positive predictions; Recall measures the proportion of actual positive cases that were correctly identified; while the F_1 score provides a balanced metric combining precision and recall.

At the **interpretative level**, the accuracy of content interpretation is assessed by examining the correctness of relationship identification, the accuracy of role attribution, and the fidelity of context preservation. This evaluation employs a scoring mechanism ranging from 1 to 10, measuring information preservation by translating extracted KGs into natural language and comparing with the source text, following the back-translation approach demonstrated in MusicBO [70].

4.4 Pipeline Implementation and Case Study

The case study examines the biographical finding aid of Andrea Costa (29 November 1851 - 19 January 1910), one of the founders of the Italian socialist movement.[5] As a political figure, Costa's biography encompasses generic life events such as birth, death, and romantic relationships, as well as specific events pertaining to political activity and interpersonal relations. These characteristics align with the expressive capabilities of RiC-O's biographical classes [78].

We develop our RIC-O-driven text-to-KG pipeline using a medium-sized LLM, opting for open-source LLM for reproducibility. Llama-3.3-70B-Instruct[6] seemed the most suitable for the task. According to Meta's official model card, it achieves 86.0 on MMLU and 92.1 in IFEval, benchmarks that evaluate the model's natural language understanding capabilities and instruction following, respectively, two factors that are crucial. The code of the whole experiment,

[5]This specific biography was selected by the domain expert as a notable example of an extensive and rich biography within the Italian National Archiving System (SAN) records, already explored in a previous work [202].

[6]`https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`

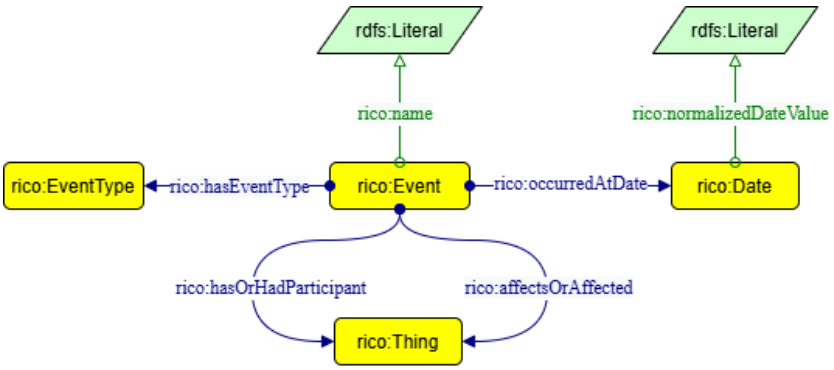


Figure 4.1: RIC-O Event core classes and properties

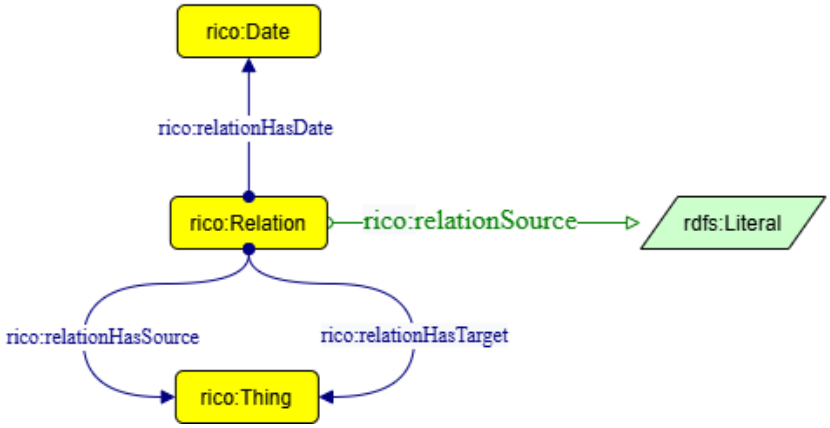


Figure 4.2: RIC-O Relation core classes and properties

alongside the results, is available as a GitHub repository.[7] Analyzing the biography alongside the domain expert, we select seven relevant events from a archive creator biography with a political career. Figure 4.1 and Figure 4.2 show the main classes and properties used within this model.

- Birth Event: mapped to `rico:Event`, connected with the subject through `rico:hasOrHadParticipant`. The binary properties are also added to the subject, `rico:hasBirthPlace`, `rico:hasBirthDate`. Family relations are established through `rico:FamilyRelation` with `rico:familyRelationType` specifying the relationship type (e.g., "parent-child").

- Family relations: represented as `rico:FamilyRelation` with prop-

[7] `https://github.com/aschimmenti/expliciting-context`

erties such as `rico:familyRelationType` (e.g., "parent-child"), `rico:hasBeginningDate`, `rico:relationHasSource`, and `rico:relationHasTarget`. Family units are modeled as `rico:Family` with `rico:hasOrHadMember` linking to family members.

- Education: modeled through `rico:PositionHoldingRelation` connecting persons to positions (e.g., student, teacher) within educational institutions (`rico:CorporateBody` with `rico:corporateBodyType`). `rico:TeachingRelation` links teachers to students, with temporal information provided by `rico:beginningDate` and `rico:endingDate`.

- Relations: implemented as `rico:AgentToAgentRelation` connecting two persons through `rico:relationConnects`, with properties such as `rico:hasBeginningDate`, `rico:hasOrHadLocation`, `rico:name`, and `rico:type` specifying the relationship type (e.g., "romantic relation"). Both agents are linked back to the relation through `rico:thigIsConnectedToRelation`.

- Political Event: represented as `rico:Activity` with `rico:hasActivityType` specifying the type of political activity. Connections to people, places, and organizations are established through various relation types: `rico:PerformanceRelation` for people's involvement, `rico:PlaceRelation` for locations. Temporal information is provided by `rico:beginningDate`.

- Production Event: modeled as `rico:Record` with `rico:hasContentOfType` specifying the type of production (e.g., "periodical publication"). Authorship is established through `rico:hasOrganicProvenance` linking to `rico:Person` or `rico:CorporateBody`. Creation date is specified with `rico:hasCreationDate`. The Production of a record is particularly important, as it can be connected with the actual record within the archive through an interface [202].

- Death Event: represented as `rico:Event` connected to the subject through `rico:hasOrHadParticipant`. Additional information such as cause of death can be included in `rico:description`. The death date is directly linked to the person through `rico:hasDeathDate`.

Although the capacity of the context window of the model (128,000 tokens for Llama 3.3) would technically allow simultaneous processing of both the input text (4,274 tokens, 16,809 characters) and the output schema, such an approach could affect the disambiguation capabilities of the model due to token density. Decomposing the extraction process into discrete steps enhances control over the pipeline. The key challenge lies in achieving optimal information density: the text must contain sufficient context to enable accurate extraction while avoiding cognitive overload that could compromise the model's performance. Therefore, we decide to process the text paragraph by paragraph, and classifying each paragraph according to one of the available event types.

4.4.1 Step 1: Text Classification

The pipeline begins with text classification to identify the types of biographical events present in the input text. The first prompt (Listing 4.1) instructs the model to classify text segments according to predefined event categories.

Listing 4.1: First prompt for text classification

```
The following text contains a snippet of the biography of {entity}.
Classify the text depending on what is being discussed. Use one or
more of the following classes and return the classification inside
a JSON. The event must be categorized independently of whether the
event is happening to {entity} or to someone mentioned in his
biography.
{classes_list}
Text: {text}
Return only a JSON array of classifications. If no proper
classification is possible, return any class with 0.0 confidence.
```

```
OUTPUT SCHEMA:
[{
  "type": "<EVENT_TYPE>",
  "confidence": 0.0-1.0,
  "reason": "<explanation>"
}]
```

Once the classification is completed, each class detected with a confidence score higher than 0.7 (an empirical threshold) is retained and used for the subsequent step.

4.4.2 Step 2: Question Answering

The second step injects a set of questions alongside the input text (Listing 4.2) with redundant instructions. Additional context and examples were incorporated into this step after multiple tests to improve model performance. The target text is presented with both preceding and following paragraphs as context.

Listing 4.2: Second prompt for question answering

```
The following text has been classified as describing a
'{event_type}' event in Andrea Costa's life.
### Context:
EVENT TYPE: {event_type}
**Previous context:** {prev_context or 'None'}
**Target text:** {text}
**Following context:** {next_context or 'None'}
### Instructions:
1. Read the questions carefully and only answer the questions with
   information relevant to the {event_type} context.
2. Read the **target text** carefully, as it contains the primary
   information you need.
3. Use the additional context (previous and following) only as
    supplementary information when the target text alone does not
```

```
   provide the full information about the {event_type} event.
4. Assume the event involves {subject} if no explicit subject is
   mentioned in the target text.
### Questions:
{questions}
### Examples:
{examples}
### Requirements:
- Provide concise, direct answers to each question in the order
  listed.
- Focus on entities, dates, and their actions.
- Avoid speculation or assumptions not supported by the provided
  text.
- Use dates in **DD/MM/YYYY** format or state the year if precise
  dates are unavailable.
- Highlight the specific relations between entities, institutions,
  and other places if any.
- Do not comment on any more information than asked.
- Keep the original language for entity labels.
- Return only the answers.

YOUR ANSWER:
```

While the second and third steps of the pipeline could be combined, separating the question answering from the JSON output reduces the workload while ensuring consistency and schema adherence. The output from this step is then sent to the final step of the pipeline.

4.4.3 Step 3: Schema-Based JSON Generation

The final prompt instructs the model to return a JSON output from the answer information following a predefined schema. Listing 4.3 shows the schema for the "Political event" class. The other event prompts can be found in the

GitHub repository.[8]

Listing 4.3: Event schema for the class "POLITICS"

```
"POLITICS": {
  "instruction": "A political event encompasses any politically
    significant action or occurrence that involves participants
    with context-dependent roles in a defined spatiotemporal
    setting. Participants must be of type PERSON, ORGANISATION,
    or GROUP. Location and temporal data are captured as separate
    elements from participant information.",
  "description": "<Detailed description of the event>",
  "properties": {
    "actions": [
      {
        "action": "<specific event or action that the entities
          suffer or cause>",
        "participants": [
          {
            "name": "<Participant's name>",
            "type": "<person/organisation/group>",
            "role": "<role of the participant in the action>"
          }
        ],
        "date": {
          "startDate": "<Start date of subject's relation>",
          "endDate": "<End date of subject's relation>"
        },
        "location": [
          {
            "label": "<Location name>",
            "description": "<Detailed description of the location>"
```

[8] https://github.com/aschimmenti/expliciting-context/blob/d6b396c647c31e91f19ac386d0a83c7d8edc1854/event_schema.json

```
          }
        ]
      }
    ]
  }
}
```

4.4.4 Post-processing and RDF Mapping

The output is additionally processed through rules to validate the JSON structure. The validated JSON output is then mapped to the classes and properties defined above. This step ensures that the extracted data adhere to a standardized semantic framework, facilitating interoperability and alignment with archival description standards.

While this modular approach introduces additional computational overhead compared to end-to-end extraction, it enables finer control over each step of the process and allows for targeted improvements where needed. However, the critical question is how effectively this pipeline performs in the specific context of archival descriptions.

4.5 Results

Evaluating structured data output from LLMs is inherently complex, requiring a combination of precision, thoroughness, and consistency. In this section, we present both quantitative and qualitative evaluations to assess the reliability of extracted data and the schema adherence of outputs. Figure 4.3 and 4.4 show two of the graphs, the first about the birth of A. Costa himself, and the second about the start of their relationship in Switzerland and the birth of Andreana, daughter of Andrea.

4.5.1 Structural Level

Evaluating structured data is a long and resource intensive process. To ensure precision and agreement between the evaluators, a simple web application was developed. It compares the extracted events to the schema (Figure 4.5). Two

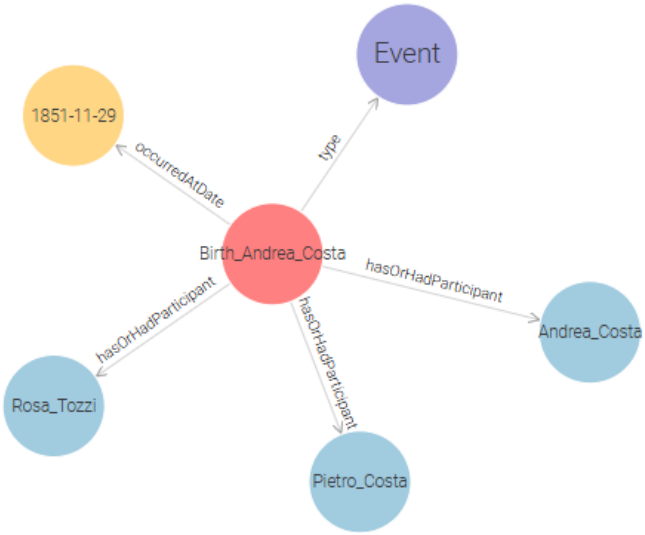


Figure 4.3: The birth of Andrea Costa

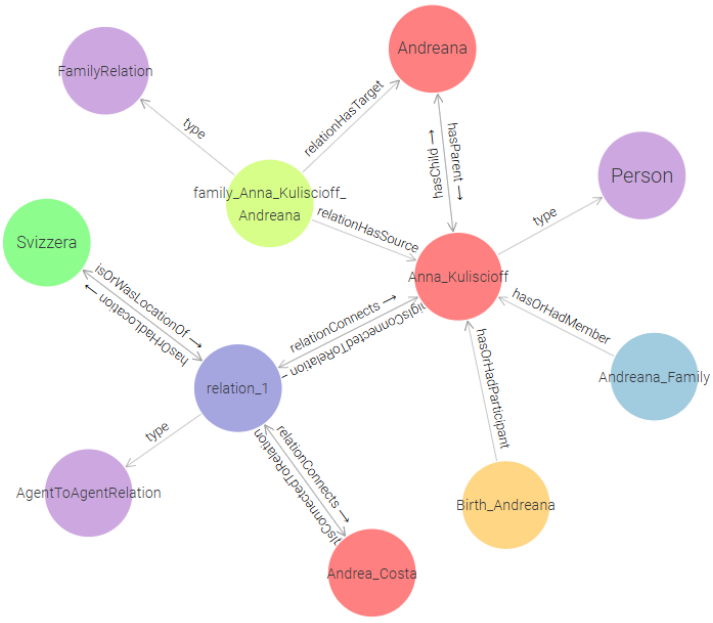


Figure 4.4: the relationship between Andrea Costa and Anna Kulischoff

check boxes track whether the output is valid and whether the classification is correct. Four radio buttons (TP, FP, FN, TN) evaluate the per-field output. Once the evaluation is completed, a report can be downloaded.

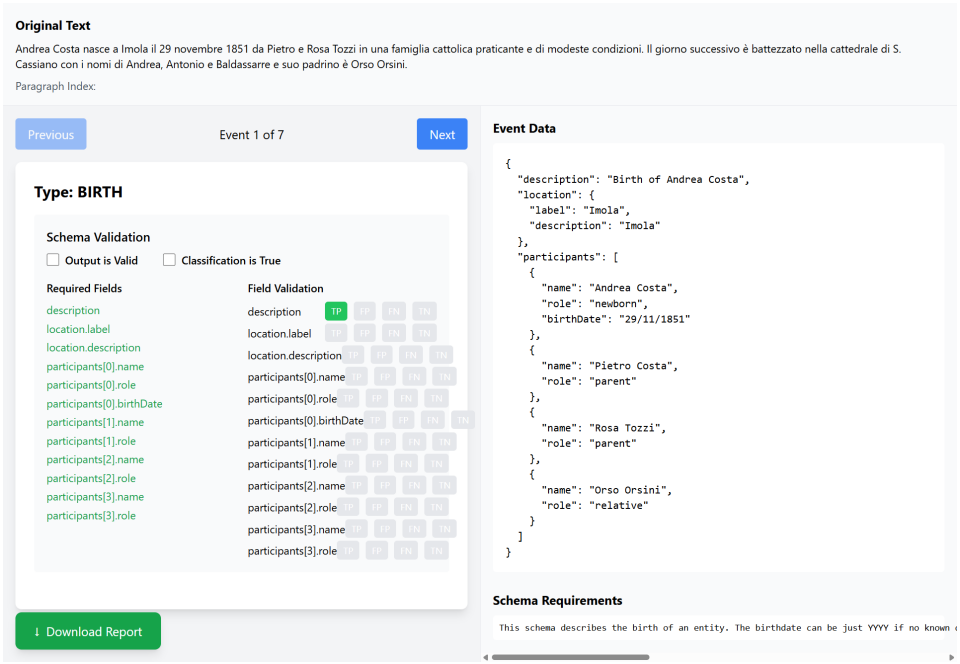


Figure 4.5: Output evaluation web application. In the example, the Birth Event of Andrea Costa is being evaluated.

One of the mentioned concerns using LLMs for structured data extraction was their capability to conform the output to a given schema. In our analysis, approximately 10% of output contained minor errors (e.g., adding a "comment" field in the schema or an additional field) or hallucinations. We classified both types as schema violations since even a 1% error rate means the JSON output becomes inconsistent and requires post-processing to ensure schema conformity. It must be noted that most of the hallucinations seemed reasonable (e.g., in one political event, the model inferred a correct date from the context but changed the key to "presumed_starting_date").

To avoid this outcome, many libraries, tools, and functions have been proposed. Current updates of LLM

Table 4.1: Valid outputs percentage

Metric	Value
Total Events	51
Schema Validity	46 (90.2%)

4.5.2 Information Level

The classification step was correct in 100% of the cases. Overlapping occurred as well, with a single event (e.g., Andrea Costa's contacts with Jules Guesde, a French politician and journalist) being structured twice, as a relationship and as a political event. We evaluated the extraction performance using both macro and micro-averaged metrics. Macro-averaging computes metrics independently for each event type and then averages them, while micro-averaging aggregates contributions across all classes by counting total true positives, false negatives, and false positives before computing metrics. This distinction is necessary because event types occur with different frequencies in biographical texts—birth and death events appear once per biography, while employment and political events may occur multiple times—and macro-averaging reveals whether the model performs consistently across all event types regardless of their prevalence [219].

The precision and recall metrics exceeded our expectations (Table 4.2). The high recall (0.982) indicates that when information is present in the input text, the model successfully extracts it, while the slightly lower precision (0.947) suggests the model occasionally produces incorrect or hallucinated information. This pattern, higher recall than precision, indicates that the model prioritizes finding all relevant information even at the cost of occasional false positives, which is often desirable in KE tasks where missing data are typically more problematic than including extraneous information that can be filtered out during post-processing. It must also be kept in mind that the inferences were performed without injecting a controlled vocabulary (apart from the entity types) for fields such as the roles.

Table 4.3 shows the micro metrics. Micro precision scores were generally

Table 4.2: Macro-average metrics

Metric	Value
Precision	0.947
Recall	0.982
F_1	0.964

above expectations except for Employment and Education events. Most of the errors were caused by the model's capability to distinguish roles and dates. It must be also noted that while the political events were generally correct, the roles were the most verbose.

Table 4.3: Micro-average precision, recall, and F_1 score per class

Class	Precision	Recall	F_1
BIRTH	0.999	0.960	0.979
DEATH	0.999	0.875	0.933
DOCUMENT	0.926	0.999	0.962
EDUCATION	0.902	0.841	0.871
EMPLOYMENT	0.840	0.971	0.901
POLITICS	0.952	0.999	0.975
RELATIONSHIP	0.999	0.955	0.977

4.5.3 Interpretative Level

Given the complexities of KE from archival records, qualitative evaluation is essential to complement quantitative metrics. While quantitative evaluation can identify structural and classification features, it does not capture the nuances of context, relationships, or roles embedded within the data. For this reason, a qualitative assessment was performed to evaluate the interpretative accuracy and contextual fidelity of extracted events.

The qualitative evaluation was conducted on an event-type basis (birth, death, education, employment, political, and record creation events), following an agreement between the evaluators. Each extracted event was compared with the original textual paragraph (associated with a "paragraph index" number) focusing on the nuances of specific event types while applying a unified scoring framework. Starting from a maximum evaluation grade of 10 (meaning that all

semantic content and contextual relationships from the source were maintained with full accuracy), points were subtracted when specific errors were identified:

- **Critical data missing**: Missing essential data (e.g., a birth date in a birth event) incurred a flat penalty of -1 point;
- **Subevent omissions**: Missing contextual details or subevents (e.g., the absence of a person's role within an event) were penalized as semi-errors, with a deduction of -0.5 points per missing element;
- **Annotation errors**: Misclassification or annotation errors (e.g., confusing the role of a person or an institution within an event) were considered significant errors, and penalized by -0.8 points;
- **Incorrect or hallucinated information**: Completely incorrect data due to hallucinations was penalized by -1 point per instance.

Each evaluation instance was documented through the indication of the paragraph index, the overall evaluation, and notes on the specific errors and their impact on the quality of the extraction. The reported scores in Table 4.4 are the result of a compromise between the two evaluators.

4.6 Discussion

The results are encouraging. The LLM-based pipeline demonstrated good performance in extracting structured information from archival descriptions, with high precision (0.947) and recall (0.982) metrics, while also providing comprehensive contextualization (8.8/10). Here we provide an example: the attempted insurrection of 1874 in Bologna and Romagna, when Costa tried to agitate multiple socialist groups alongside Antonio Cornacchia, with the anticipated presence of Michail Bakunin.

Input: "1874, a year of severe economic crisis, marked by widespread popular exasperation and numerous protests especially against the grain tax, was chosen by Italian internationalists for their first insurrectionary attempt.

Table 4.4: Qualitative evaluation results per event class

Class	Mean Score	Support	Performance	Errors
Birth	8.8	7	Generally accurate	Missing dates with multiple concurrent birth events
Death	10.0	2	Well structured, dates and locations correct	—
Relationship	9.5	4	Well structured, dates and locations correct	Overstated roles (e.g., "alliance" for opinion agreement)
Education	8.4	3	Generally accurate	Errors in role descriptions and dates with concurrent events
Employment	8.8	6	Generally accurate	Errors in role descriptions and dates with concurrent events
Politics	8.3	26	Good grasp of context and events	Roles sometimes too generic (e.g., "His followers" instead of "Andrea Costa's Followers")
Document	8.1	3	Generally accurate	Missing metadata (date, location)
Total	**8.8**	**51**	Good grasp of context, succession, entities	Issues with role assignment and implicit dates/locations

Costa was the main organizer of the movement, which was supposed to have Bologna as its epicentre and see the participation of Bakunin in person. However, inadequately organized, lacking the expected popular support, and preventively monitored by the police, the attempt failed miserably: Costa was arrested even before the attempt, and the small column of Romagna's Internationalists marching on Bologna under the command of Antonio Cornacchia was captured without a fight, while Bakunin managed to slip away, incognito as he had arrived."

Listing 4.4: JSON output of the model from the text (simplified)

```
Action: Attempted insurrection

    Time Period:

      - Begin: 1874
```

```
    - End: -
  Locations:
    - Bologna (intended epicenter of the insurrection)
    - Romagna (origin of the Romagnol internationalists)
  Key Participants:
    - Andrea Costa
      Role: Primary organizer
      Type: Individual
    - Antonio Cornacchia
      Role: Commander of the Romagnol internationalists
      Type: Individual
    - Bakunin
      Role: Expected participant
      Type: Individual
    - Italian internationalists
      Role: Participants
      Type: Group
    - The Police
      Role: Opponent
      Type: Organisation
```

The JSON outputs were then converted through a simple rule-based script into Turtle syntax, using RiC-O. Listing 4.5 shows part of the Turtle-format version of the same example:

Listing 4.5: Part of the RDF version of the JSON in Listing 4.4

```
@prefix ex: <http://example.org/#> .
@prefix rico: <https://www.ica.org/standards/RiC/ontology#> .

ex:activity_failed_insurrection_attempt_by_italian_internazionalists_9
  a rico:Activity ;
  rico:hasActivityType ex:activity_type_attempted_insurrection ;
  rico:name "Failed insurrection attempt by Italian
```

```
    internationalists" ;
  rico:relationHasTarget ex:andrea_costa .

ex:andrea_costa
  a rico:Person ;
  rico:name "Andrea Costa" .

ex:andrea_costa_failed_insurrection_attempt_by_italian_internazionalists_9
  a rico:PerformanceRelation ;
  rico:description "principal organizer" ;
  rico:relationHasSource ex:andrea_costa .
```

As additional evaluation, the JSON output (Listing 4.4) can be converted by an LLM into a paragraph for readability during the qualitative evaluation step. The JSON content, elaborated by Llama 3.3 70B, provides the following reconstructed text:

> In 1874, an insurrectional plot began to take shape in Italy, concentrated mainly in Bologna with connections in the Romagna region. The main organiser was Andrea Costa, who worked together with Antonio Cornacchia, commander of the Romagna internationalists. The group of Italian internationalists acted as both organiser and participant, and Bakunin was expected to join as a participant and leader in the insurrection attempt.

While factually correct, an in-depth analysis of the output raises additional challenges. Event representation is a known challenge, especially for complex historical events through semantic data structures. In this case, the problem is related to the definition and labelling of participant roles. The current implementation shows inconsistencies in labelling, moving between broad categorizations and highly detailed descriptions that embed temporal or conditional information directly in the role labels. The 1874 insurrection attempt provides a clear example: Bakunin is labeled an expected participant, which

creates two key problems. First, it places factual and anticipated participation on the same semantic level. Second, it affects data retrieval accuracy: query for participants in the 1874 insurrection would include both actual participants and those who were merely expected to take part, reducing the truth value of the query. This inconsistency highlights the need for controlled vocabularies that can be suggested to the LLM to standardize role descriptions while maintaining the semantic richness of the original text.

The qualitative evaluation, while generally reporting promising results, also revealed concerns about knowledge loss during the extraction process. While the LLM-based approach showed significant accuracy in identifying and structuring explicit information, some nuanced contextual information expressed in the original narrative form may not be captured by the event schema. To assess the relevance of the lost information from the functional point of view, the engagement of archivists and users is mandatory.

Several directions for future research emerge from these findings:

- LLMs can bootstrap text-to-KG pipelines but the design still requires extensive unit testing (e.g., for the vocabulary alignment). It is necessary also to test on a broader corpus of case studies that should include a significant variety of biographical profiles, also pertaining to different eras and cultures;

- In this proof of concept, the evaluation was performed without a reference ground truth, being directly evaluated by two humans. In future and larger works, it is necessary to annotate enough data to have an explicit reference ground truth;

- The integration of additional event types and more expressive schemas can better represent the complex relationships available in archival descriptions, particularly focusing on temporal and contextual dimensions. Integrating Frame Semantics, partially following FRED's approach [71], could help create more dynamic templates while retaining flexibility;

- The exploration of hybrid approaches that combine LLM capabilities to restrict outputs and evaluate automatically the distance between the output and the source text.

However, some limitations need to be acknowledged. First, the document's format is exceptionally precise in describing events (both from Costa's personal and political life) individually, with clear paragraph-level boundaries that align well with our one-paragraph-at-a-time processing approach. This structural characteristic of the Costa finding aid may have contributed to the pipeline's strong performance, as each paragraph functioned as a relatively self-contained biographical event description. This raises questions about generalization to finding aids with different organizational structures, such as those employing more narrative-driven or chronologically integrated formats where events are not as clearly delineated. Second, the absence of systematic baseline comparison limits our ability to attribute performance gains to specific pipeline components. While the results demonstrate that the overall architecture performs well, we cannot determine the relative contributions of individual design choices—such as the three-step modular approach versus potential end-to-end extraction, the specific prompt engineering strategies employed, or the choice of Llama-3.3-70B versus alternative models. An ablation study examining variations in model size, prompt complexity, and architectural decisions would provide clearer insights into which components are essential for achieving similar performance levels and which could potentially be simplified without significant quality degradation. Third, the evaluation was conducted on a single finding aid selected for its richness and complexity rather than its representativeness. While this choice enabled thorough qualitative assessment and provided proof of feasibility, it prevents conclusions about how the pipeline would perform across the diverse range of biographical profiles, archival description practices, and historical periods present in archival collections. The Costa biography represents a specific type of archival description, and performance may vary significantly for finding aids describing different types of

creators (e.g., artists, private individuals, corporate bodies) or materials from different cultural and temporal contexts.

4.7 Conclusion

This chapter proposed an ontology-driven text-to-KG pipeline applied to archival finding aids, demonstrating how LLMs can be integrated into multiple stages of the extraction process. The pipeline successfully translated biographical narratives from the Andrea Costa finding aid into RDF triples aligned with RiC-O, achieving high precision (0.947) and recall (0.982) metrics alongside strong interpretative fidelity (8.8/10). These results indicate that LLMs can effectively bootstrap text-to-KG systems for archival descriptions without requiring extensive domain-specific training data for each subtask.

The proof of concept validated several capabilities central to LLM-based KE in archival contexts. The model demonstrated robust performance in classifying biographical event types, extracting structured information from narrative text, and adhering to predefined schemas with 90% validity. One of the examples we discussed in the previous chapters, i.e. inferring from context the full name of an entity, was successfully executed. The qualitative evaluation revealed that the pipeline maintained semantic accuracy across diverse event categories, from basic life events to complex political activities involving multiple participants and temporal dimensions. The back-translation approach confirmed that extracted information could be reconstituted into coherent narrative form, preserving essential contextual relationships from the source text.

However, the evaluation also identified critical challenges that must be addressed in future implementations. Schema violations, while infrequent, highlight the need for robust validation mechanisms in production systems. The inconsistent handling of participant roles—particularly the conflation of actual and anticipated participation—demonstrates that controlled vocabularies and ontological constraints must be more tightly integrated into the extraction process. The loss of nuanced contextual information during transformation from

narrative to structured format raises questions about the appropriate balance between standardization and expressiveness in archival KE systems.

The modular architecture adopted in this proof of concept, keeping separate classification, question answering, and schema generation into discrete steps, proved essential for maintaining control over the extraction process and enabling targeted improvements. This decomposition allowed for iterative refinement of individual components while preserving overall pipeline coherence. The approach also facilitated transparent evaluation, as each step could be assessed independently before examining end-to-end performance.

The findings point toward several methodological considerations for future text-to-KG implementations in archival and Cultural Heritage (CH) contexts. First, the extensive prompt engineering and iterative testing required to achieve satisfactory results suggests that developing such pipelines demands systematic approaches to design and validation. Second, the need to balance schema expressiveness with extraction accuracy indicates that ontological analysis must inform every stage of pipeline development, from event categorization to output validation. Third, the interpretative nature of the extraction process necessitates explicit provenance documentation, distinguishing computationally derived assertions from authoritative archival descriptions.

While our experiments demonstrated feasibility on a single, carefully selected finding aid, broader adoption requires addressing questions of generalizability, scalability, and institutional integration. The pipeline must be tested across diverse biographical profiles spanning different historical periods, cultural contexts, and archival traditions. The evaluation framework, while sufficient for focused case studies, needs extension to support larger-scale assessments with explicit ground truth annotations. Beyond technical validation, the approach requires engagement with archival practitioners and end users to assess whether the extracted structured data adequately serves research and access needs without sacrificing the interpretative richness that characterizes professional archival description.

These considerations reveal the need for a comprehensive methodological framework that extends beyond individual proof of concepts to address the systematic design, implementation, and evaluation of text-to-KG pipelines in CH contexts. Such a methodology must account for the iterative nature of pipeline development, the integration of domain expertise throughout the process, and the validation requirements specific to generative KE approaches. The following chapter presents such a framework, building upon the insights gained from this archival case study to propose a generalizable approach for developing text-to-KG systems across diverse CH applications.

Chapter 5

Knowledge Extraction in the Curation life cycle: The ATR4CH Methodology

> "The challenge is to shift humanistic study from attention to the effects of technology [...] to a humanistically informed theory of the making of technology (a humanistic computing at the level of design, modeling of information architecture, data types, interface, and protocols)."
>
> — Drucker, *"Humanistic Theory and Digital Scholarship"* (2012)

This chapter introduces *Adaptive Text-to-RDF for Cultural Heritage* (ATR4CH), an iterative methodology for designing text-to-KG pipelines with a focus on annotation, Large Language Models (LLMs) integration, and RDF. The methodology stems from challenges and insights gained through two previous experiments on Knowledge Graph (KG) generation from text: the proof of concept of archival finding aids presented in Chapter 4 and an additional preliminary case study on scholarly debates on forged documents [179]. The methodology, inspired by Cultural Heritage (CH) curatorial practices, can theoretically be applied to other domains with similar requirements. Section 5.1 discusses the objective of the methodology, exploring the role of KE within digital curation, and how pipeline design decisions shape what knowledge becomes computationally accessible. Section 5.2 discusses the methodological

frameworks that informed the development of ATR4CH. Section 5.3 presents the methodology itself[1].

5.1 Background

Digital curation encompasses the activities required to maintain, preserve, and add value to digital resources throughout their life cycle [92]. Testoni [194] characterizes the digital curator as a "producer of meaning"—an agent who facilitates the creation of digital objects through systematic intervention: selecting resources, determining their description and ensuring their preservation by applying semantic layers that follow particular perspectives [201]. Constrained by community practices and standards, this curation process nevertheless constitutes an interpretative act. Following Stachowiak's model theory [185], ontologies embody a *reduction property*: they capture only those attributes their creators deem relevant, functioning for particular subjects, within particular contexts, and restricted to particular operations [155]. Knowledge representation models encode interpretative choices made during ontology development, reflecting particular conceptualizations of domains [84, 147]. These modeling decisions and the resulting standards carry epistemic consequences for knowledge that is deemed important to be computationally accessible. Within the curatorial life cycle, KE is an integral component. Although often characterized as a set of merely technical activities[2], KE "inherits" the interpretative position embedded in the ontological models that guide it and of the tools used to accomplish it. This aspect of KE manifests through multiple interconnected dimensions. At the ontological level, representational choices determine what categories of knowledge merit formalization. At the technical level, the limitations of the extraction tool impose constraints on what phenomena can be reliably captured. NLP tools trained on commercially-driven, well-

[1]This chapter is an extension of prior work by the author: Schimmenti, A., Pasqual, V., Vitali, F., & van Erp, M., *Knowledge Graphs Generation from Cultural Heritage Texts: Combining LLMs and Ontological Engineering for Scholarly Debates* (2025, submitted to *Journal of Documentation*), currently under review.

[2]e.g. `https://vocabs.dariah.eu/tadirah/`

resourced datasets serve domains with abundant training data while struggling with specialized CH contexts. These two dimensions (ontological representation and technical extractability) exist in bidirectional tension. The ontology drives the requirements of KE tools by defining target knowledge structures, while these tools' capabilities may constrain what can be represented feasibly from texts. When an ontology addresses domains lacking dedicated NLP resources, successful text-to-KG implementation requires accommodating and sometimes compromising representational choices. The risk is that tool availability rather than scholarly requirements determines what knowledge receives computational representation, thereby compounding existing biases in archival and documentary practices. Automatic KE practices are not neutral elements of the curatorial process: they constitute active agents that align with particular interests and necessities and must be carefully selected, adapted, and evaluated. The availability of extraction tools creates a form of *data availability bias*[3]: when curators recognize that existing NLP resources cannot adequately serve ontologies representing less-studied domains, they may deprioritize these domains in project planning. The absence of suitable tools signals to researchers that certain narratives remain computationally inaccessible, potentially discouraging proposals for long-tail domains or edge cases. One counterexample from the trends in Chapter 2 demonstrates how this bias can be addressed: the representation of individuals such as colonial subjects who were not presented in the sources with a name, ending up not represented in traditional archival facets, could not be extracted by standard NER tools, requiring specific methodological adaptations [130]. This observation highlights two methodological needs: designing approaches that systematically prevent entities, concepts and relations from being excluded due to tool limitations, while maintaining transparency about what extraction capabilities can realistically be developed. ATR4CH addresses this curatorial complexity through several methodological commitments. Rather than developing ontologies within

[3] `https://combattingbias.huygens.knaw.nl/bias/types/availability/`

the pipeline design process, the methodology takes ontological models as foundational inputs, adapting annotation schemas and extraction architectures to align with predetermined representational frameworks. Evaluation of the output becomes central, necessitating explicit generation of ground truth (annotated data) to test the pipeline rather than relying on post-hoc evaluation alone. The methodology positions this process not as mere data preparation, but as intensive interpretative engagement that develops domain expertise, revealing both extraction challenges and potential ontological misalignments among the data, the ontology and available NLP tools. LLMs address data scarcity in underrepresented domains through few-shot and zero-shot learning [27], as discussed in the previous chapter. ATR4CH explicitly incorporates LLMs within text-to-KG pipelines, where annotation serves dual purposes: as ground truth for evaluating ready-to-use models and as training data enabling supervised fine-tuning when computational resources permit.

5.2 Theoretical Foundations

ATR4CH builds upon established practices in knowledge engineering and ontology development. The methodology draws from agile ontology engineering approaches documented in the literature, particularly METHONTOLOGY [66], SAMOD [151] and eXtreme Design (XD) [158] among the many ontology design methodologies [102, 8]. These methodologies share several characteristics that inform ATR4CH's design:

- **Competency Question centrality**: Design decisions, from pattern selection to validation, are driven by CQs that formalize user requirements
- **Test-driven development**: Unit tests derived from CQs validate ontology modules before integration
- **Modular design**: Content Ontology Design Patterns function as reusable modules addressing recurrent modeling challenges [69]
- **Iterative refinement**: Development proceeds through repeated cycles

rather than sequential phases through evolving prototypes

- **Domain expert involvement**: Specialists continuously validate design decisions throughout the development process

Analysis of the projects surveyed in Chapter 2 reveals a recurring pattern in CH implementations: ontological modeling precedes the design of KE strategies. Odeuropa, for instance, developed its ontology framework before designing extraction approaches [122]. This sequence presents two methodological options for coordinating ontology development and knowledge extraction:

- A unified methodology that coordinates both ontology engineering and knowledge extraction

- A methodology focused exclusively on knowledge extraction design, assuming a pre-existing ontological framework

The projects analyzed in Chapter 2 demonstrate both scenarios: some projects develop ontologies concurrently with extraction systems, while others enrich existing models through automatic population. ATR4CH adopts the second approach to accommodate projects in both cases while maintaining methodological coherence. ATR4CH synthesizes practices observed across the text-to-KG projects documented in Chapter 2 into a systematic methodology adaptable to CH contexts. The approach maintains alignment with established ontology engineering principles to reduce adoption barriers for institutions already employing these methodologies. The coordination of annotation schemas with target ontologies follows patterns established in projects such as Odeuropa, where frame-based annotation schemes were developed in parallel with the Odeuropa ontology[4] to extract smell event mentions from historical texts [122]. This approach treats annotation development as designing intermediate representations that bridge discourse structure and RDF formalization. ATR4CH establishes three evaluation perspectives for generated KGs:

[4] `https://data.odeuropa.eu/ontology/`

component-level metrics (precision, recall, F_1-scores) assess individual extraction tasks; Competency Question (CQ) answering evaluates whether the KG supports the information requirements that motivated its construction; semantic fidelity assessment through back-translation measures representational completeness.

5.3 The ATR4CH Methodology

ATR4CH assumes the presence of the following inputs:

- A corpus of documents containing knowledge to be extracted;
- A target ontology defining the conceptual framework for representation;
- A set of CQs (either derived from the ontology development process or specifically elaborated for the text-to-KG pipeline).

Before describing the methodology workflow, the following definitions introduce key concepts that structure ATR4CH's approach to incremental text-to-KG development.

- **Core Ontological Pattern (COP)**: Essential KG pattern representing central ontological nodes and relationships that are both present in the corpus as extractable information and necessary for addressing the CQs. A COP can be defined operationally as the set of KG structures that satisfy a SPARQL query corresponding to a CQ. Not all ontologies are designed using explicit Content Ontology Design Patterns [158]; however, every ontology embodies recurrent structural patterns that can be identified through its intended use cases and CQs. *Example*: In CIDOC CRM, a common COP is the Creation Event pattern: "Artwork Y was created by Artist X at Time T using Technique M," corresponding to CQs such as "Who created this artwork?" The SPARQL query retrieving creator, time, and technique information defines the KG structure that extraction must produce;

- **Pilot Corpus**: Small set of representative documents (3-5 documents) serving as a development sandbox for annotation and KE exploration, selected for qualitative representativeness. *Example*: For extracting artist biographies, a Pilot Corpus might include a museum wall text (brief, formal), an exhibition catalog entry (structured, technical), and a newspaper article (narrative, contextual);

- **Minimal Working Annotation (MWA)**: Annotation schema that emerges from COP-specific iterations, capturing essential knowledge structures necessary for extracting identified COPs. The MWA aggregates annotation schemas developed for individual COPs during pilot development, following the principle of iterativity. *Example*: For the Creation Event COP, the MWA might include annotation layers for artist mentions (spans + Wikidata IDs), temporal expressions (spans + normalized dates), technique terms (spans + Getty AAT concepts), and creation relationships connecting these elements;

- **Production Annotation Model**: Comprehensive annotation schema evolved from the MWA, suitable for creating ground truth on test data. May include additional elements such as coreference chains spanning multiple COPs, disambiguation tags, or confidence indicators. *Example*: Extending the MWA, the production model might add coreference chains linking "Picasso," "he," and "the artist"; disambiguation tags distinguishing "Blue Period" (artistic movement) from "blue" (color); confidence scores for uncertain attributions;

- **Ground Truth**: Manually annotated test dataset created using the production annotation model, serving as gold standard for systematic evaluation. This dataset is separate from the Pilot Corpus used during development. *Example*: 25 artist biography documents fully annotated with all COPs, used to compute precision, recall, and F_1-scores for each extraction component and to evaluate whether the pipeline answers the

original CQs;

- **Rehydration**: We introduce the term *rehydration* to generalize the back-translation approach employed in MusicBO [70]. In traditional back-translation, the goal is to convert a structured representation (e.g., a KG) back into natural language and then compare the generated text with the original source to assess fidelity. This approach assumes that the KG fully encodes the information present in the source text, so that the reconstructed and original texts express the same content.

  In our setting, however, the KG may capture only a subset of the source document—specifically, the information relevant to the identified COPs. In such cases, the notion of "translation" becomes misleading, since the KG and the source text are not semantically equivalent but stand in a part–whole relation. To better describe this process, we adopt the term *rehydration*, which denotes the reconstruction of natural language text from the information contained in the KG, regardless of whether the KG represents complete or partial source content. When the KG encodes the full information of the source, standard overlap-based evaluation metrics developed for machine translation—such as BLEU [146], METEOR [15], or CHRF++ [157]—remain applicable. However, when the KG represents only partial content, these metrics introduce bias by penalizing the absence of information that was considered "noise". In such cases, rehydration uses metrics suited for partial content comparison, such as summarization metrics or semantic similarity measures like G-EVAL [123] and BARTScore [224], which leverage LLMs or pretrained language models to assess the fidelity of the reconstructed text to the extracted information without penalizing omissions. Following [91], these approaches offer a fairer measure of semantic correspondence in selective extraction scenarios. Rehydration can be operationalized through template-based systems that systematically traverse KG structures or via LLMs that generate fluent text from structured inputs [73]. This

process can also facilitate expert validation by providing interpretable natural language renderings of structured knowledge [148].

Figure 5.1 presents the workflow structure of ATR4CH. The methodology follows a Closed text-to-KG paradigm, accommodating projects with varying resource constraints and corpus sizes. Drawing from agile software development practices as instantiated in XD [158], ATR4CH organizes work around tasks rather than rigid sequential phases. Tasks can be executed iteratively and retraced based on evaluation results, allowing the methodology to adapt to emerging requirements and technical constraints discovered during development.

The ATR4CH methodology adopts an *incremental, pattern-by-pattern development strategy* to implement ontology-grounded information extraction. Rather than attempting to process or map the entire ontology at once, ATR4CH iteratively focuses on one Core Ontological Pattern (COP) at a time, identified through analysis of the target ontology and its competency questions. Each COP progresses through the complete development cycle—from annotation schema design, through RDF mapping validation, to the implementation of automated extraction modules—before moving to the next pattern. This focused approach enables systematic validation and refinement at each iteration, ensuring that both the ontology alignment and extraction procedures remain empirically grounded.

After several COP-specific iterations, the methodology transitions from the *Pilot Corpus phase* to full-scale corpus processing. The Minimal Working Annotation (MWA), developed during these early iterations, is consolidated into a production-ready annotation model that supports comprehensive ground truth creation and pipeline evaluation. This transition reflects ATR4CH's dual commitment to agile development and rigorous assessment: intensive iteration on representative examples followed by systematic validation on unseen test data.

The overall workflow, illustrated in Figure 5.1, is organized into five in-

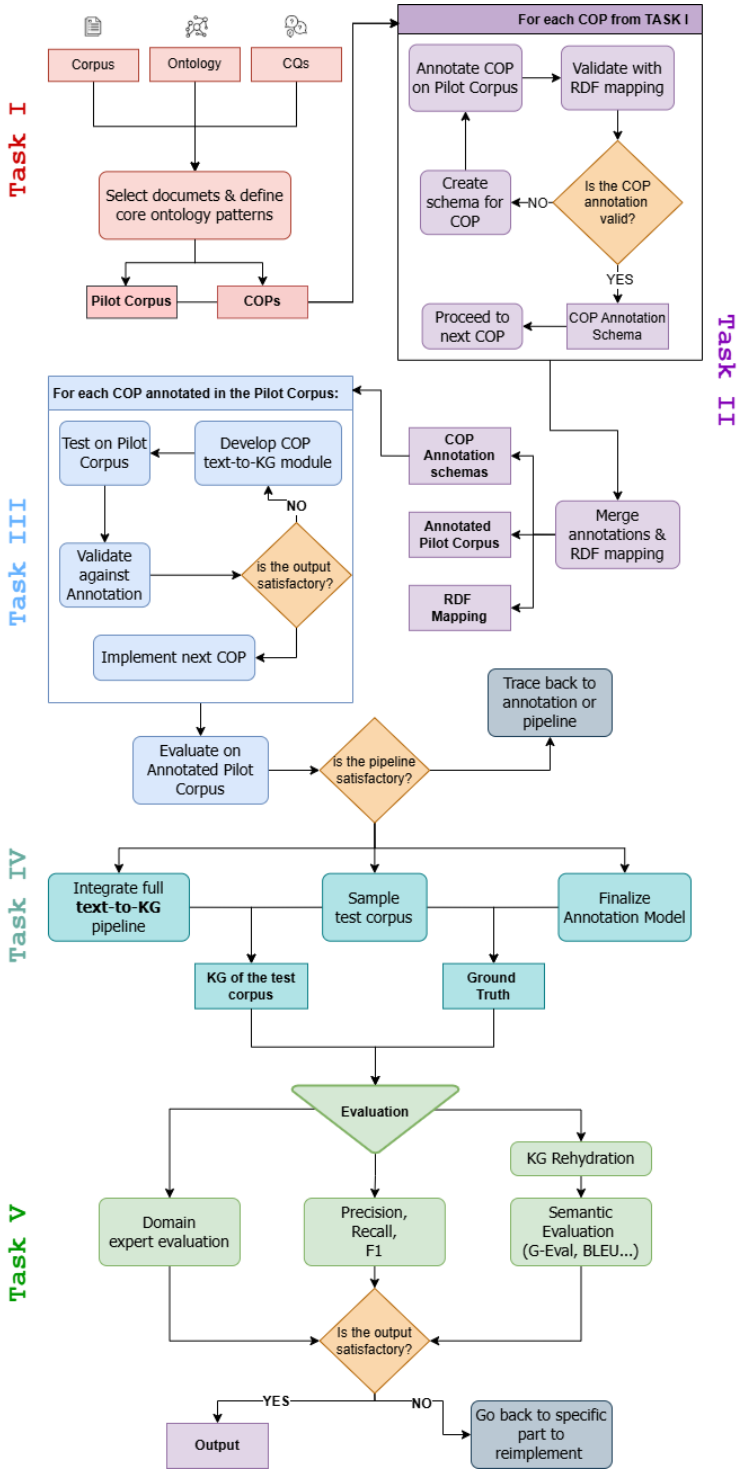


Figure 5.1: Workflow of the ATR4CH methodology showing the iterative task structure.

terrelated tasks. While presented in numbered order for clarity, these tasks are modular components within an iterative process rather than sequential

waterfall stages. Each color in the figure corresponds to one methodological task, detailed in Sections 5.3.1–5.3.5. The methodology explicitly supports returning to earlier stages when evaluation results suggest the need for revision, thus maintaining flexibility and empirical responsiveness throughout the development process.

- **Task I** (red in Figure 5.1; see Section 5.3.1) — *Foundational Analysis and Design*: establishes the conceptual and empirical foundation by analyzing the corpus and ontology to identify Core Ontological Patterns (COPs) and select the Pilot Corpus.

- **Task II** (purple in Figure 5.1; see Section 5.3.2) — *Minimal Working Annotation Development*: designs, applies, and validates COP-specific annotation schemas on the Pilot Corpus, producing the Minimal Working Annotation (MWA).

- **Task III** (blue in Figure 5.1; see Section 5.3.3) — *Pipeline Architecture Development*: implements and tests automated extraction modules for each COP, validating them against manual annotations.

- **Task IV** (teal in Figure 5.1; see Section 5.3.4) — *Integration and Refinement*: consolidates COP-specific modules into a unified text-to-KG pipeline and evolves the MWA into a production annotation model.

- **Task V** (green in Figure 5.1; see Section 5.3.5) — *Knowledge Extraction and Evaluation*: applies the integrated pipeline to the test corpus, performs quantitative and semantic evaluation (including rehydration), and supports domain expert validation.

Domain expertise is required throughout the methodology. The expertise is essential for: (1) corpus and ontology analysis (Task I), where domain knowledge guides COP identification and corpus selection; (2) annotation schema design (Task II), where specialists determine what information warrants extraction and how it maps to ontological structures; (3) validation and refine-

ment throughout all tasks, where experts assess whether extracted KGs adequately represent source semantics and satisfy research requirements; and (4) final evaluation (Task V), where domain specialists verify extraction quality.

5.3.1 Foundational Analysis and Design (Task I)

Task I establishes foundational understanding of source materials by analyzing both corpus and ontology to identify COPs for KE. This analysis addresses data sparseness problems common in tasks on unstructured texts. The task comprises four interconnected activities:

- **Corpus Analysis**: This ontology-dependent analysis examines how knowledge manifests throughout textual discourse, including linguistic patterns, discourse structures, and representational strategies. The analysis identifies which document sections contain extraction-relevant information versus tangential content. Key challenges include implicit mentions requiring contextual inference, long-distance dependencies where KG components are separated by substantial text spans, nested entities discussed through relational structures, and ambiguous references using nicknames or figures of speech;

- **Ontology Analysis**: This parallel activity assesses which parts of the target ontology can be populated from source documents, examining alignment between the ontology's conceptual framework and available textual information. It identifies which ontological classes and properties have sufficient textual evidence, which relationships are present or can be inferred from corpus discourse patterns, and which elements may need omission due to lack of textual support. The aim is determining *what* data is present in sources and what requires integration from elsewhere, rather than immediately addressing *how* to extract. CQs guide prioritization of ontological coverage based on research requirements;

- **COPs Identification**: Based on corpus and ontology analyses, this activity identifies the COPs required to answer the CQs. The identifica-

tion process involves: (1) assessing alignment between CQs, ontological structures, and available textual content, (2) identifying patterns with sufficient textual evidence for reliable extraction, (3) prioritizing based on extractability feasibility and CQ relevance, and (4) selecting a manageable subset that forms the semantic backbone for KE. These COPs will be processed incrementally, one at a time, through the subsequent tasks;

- **Pilot Corpus Selection**: Selection of the Pilot Corpus ensures coverage of various linguistic patterns, discourse structures, and diverse manifestations of the target COPs while remaining manageable for intensive manual development work. The corpus size (3-5 documents) depends on document length and complexity of information manifestation patterns detected during corpus analysis. This same set of documents will be used iteratively for developing and validating extraction pipelines for each COP.

5.3.2 Minimal Working Annotation Development (Task II)

Task II develops annotation schemas incrementally, processing one COP at a time rather than attempting to annotate all patterns simultaneously. For each COP identified in Task 5.3.1, an annotation schema is developed, applied to the Pilot Corpus, and validated through RDF mapping before proceeding to the next pattern. This task produces the MWA, which serves as the target schema for automated extraction.

Pattern-by-Pattern Schema Development: Development proceeds iteratively through the identified COPs. For each pattern, an annotation schema is designed that captures the essential knowledge structures while remaining practical for both manual annotation and automated extraction. The schema design accounts for diverse ways that knowledge manifests in the corpus, including both explicit textual mentions and information requiring infer-

ence or contextualization.

The annotation schema for each COP should include only elements necessary for extracting that specific pattern, avoiding over-annotation in the first iterations that complicates extraction without contributing to answering the CQs. If the pattern requires complex semantic structures beyond simple triple patterns, the annotation schema should include appropriate mechanisms for representing these relationships in ways reliably mappable to RDF (e.g., if the ontology relies on Named Graphs).

Knowledge Base Integration Strategy: Knowledge base integration enables consistent entity identification and vocabulary alignment between textual mentions and the target ontology. Since KGs typically involve individuals or controlled vocabularies defined within or referenced by the ontology, annotators need access to these resources to ensure textual references link to correct ontological entities. Without this integration, the same real-world entity might be annotated inconsistently across documents, preventing proper aggregation and reasoning in the final KG.

This integration, whether through building local vocabularies or leveraging external resources such as Wikidata or DBpedia, must be designed early to establish clear protocols for entity linking and vocabulary alignment that will guide both manual annotation and automated extraction in Task 5.3.3. The choice between local and external knowledge bases depends on domain coverage, data quality requirements, and specific entity types required by the COPs. Early integration ensures the annotation schema consistently handles entity disambiguation, coreference resolution, and terminological standardization throughout development.

Annotation Paradigm: Annotation should follow established practices from corpus linguistics and NLP. When resources permit, multiple annotators should annotate the same documents to enable measurement of inter-annotator agreement using metrics such as Cohen's kappa [30, 39] or Krippendorff's alpha [112]. This assessment identifies ambiguous annotation categories and

reveals where guidelines require clarification. In resource-constrained settings, a single experienced annotator may suffice, but annotation guidelines should be thoroughly documented to ensure reproducibility. The annotation process should be iterative: initial guidelines are refined based on encountered edge cases and difficult decisions during pilot annotation.

Iterative Development Process for Each COP: For each identified COP, development follows a systematic cycle ensuring the annotation schema produces RDF structures satisfying that specific pattern:

1. **Schema Design**: Develop annotation layers for the current COP, incorporating knowledge base integration protocols through tagsets, controlled vocabularies, and standardized terminologies aligning with the target ontology;

2. **Pilot Corpus Annotation**: Annotate the entire Pilot Corpus for the current COP using the schema iteration to identify potential gaps, inconsistencies, or practical bottlenecks;

3. **Mapping Validation**: Conduct RDF mapping from annotated data to RDF format, testing whether resulting KGs satisfy ontological constraints for this COP and adequately represent the semantic content of source documents. This mapping serves as a unit test for the annotation schema, validating that individual annotation patterns correctly transform to valid RDF;

4. **Schema Refinement**: Refine the annotation model based on issues identified during mapping validation, returning to previous activities as necessary;

5. **Pipeline Development for Current COP**: Once the annotation schema for the current COP has been validated through successful RDF mapping, proceed to Task 5.3.3 to develop automated extraction for this pattern using the same Pilot Corpus documents. Only after completing

pipeline development and validation for the current COP should development proceed to the next pattern.

The MWA emerges as the aggregation of annotation schemas developed for individual COPs.

5.3.3 Pipeline Architecture Development (Task III)

Task III designs and implements computational tools to automatically extract KGs from text, processing one COP at a time in alignment with the pattern-by-pattern approach established in Task 5.3.2. Rather than building the complete extraction pipeline before testing, this task develops and validates extraction capabilities incrementally for each COP using the same Pilot Corpus documents that were annotated in the previous task.

Task Decomposition: the pipeline is developed around each COP, going in hierarchal order, from the ones with higher priority to the rest. For the COP currently being processed, the extraction pipeline targets the relative annotation schema developed in Task 5.3.2. Tool choice aligns with available resources and data characteristics. For instance, regarding pre-trained LLMs vs supervised approaches:

- **Low data, low resources**: API-based LLMs with few-shot prompting and rule-based entity linking;
- **Moderate data, moderate resources**: Hybrid approaches combining pre-trained models with domain-specific fine-tuning;
- **Large data, extensive resources**: Custom model training and ensemble methods;
- **Large data, low resources**: Structured pipeline approaches leveraging smaller models with knowledge distillation.

LLM-based approaches may use structured output generation through JSON schemas [175, 159] and In-Context Learning strategies [27, 136], com-

bined with specialized NER tools [44] for precise span identification when character-level accuracy is critical.

Pipeline Implementation: Development targets the annotation schema for the current COP, integrating knowledge base resources and vocabulary standardization protocols through prompt integration or e.g. Retrieval-Augmented Generation [116]. Initial implementation focuses on basic functionality for the current COP before optimization. This task produces extraction components capable of processing raw text and generating structured outputs following the annotation schema for the specific pattern being developed.

Immediate Validation: Once extraction for the current COP has been implemented, the pipeline is tested on the Pilot Corpus and results are compared against the manual annotations created in Task 5.3.2. This immediate validation cycle enables rapid identification of extraction bottlenecks or misalignments before proceeding to the next COP. If validation reveals issues, return to Task 5.3.2 for annotation schema refinement, or to Task 5.3.1 if fundamental ontological misalignments are discovered.

Only after successfully validating extraction for the current COP should development proceed to the next pattern, returning to Task 5.3.2 to develop its annotation schema. This cycle continues until extraction pipelines have been developed and validated for all identified COPs using the Pilot Corpus.

5.3.4 Integration and Refinement (Task IV)

Task IV integrates the individual COP-specific extraction components developed in Task 5.3.3 into a unified pipeline, refining the system through end-to-end testing to achieve production-ready status. This task represents a critical transition point: the MWA that emerged from processing individual COPs must now be consolidated into a production annotation model suitable for full corpus processing.

Pipeline Integration: The modular extraction components developed for individual COPs are integrated into a unified text-to-KG pipeline. Integration addresses dependencies between patterns, ensures consistent entity

resolution across components, and optimizes the overall processing architecture. The integrated pipeline processes documents from raw text to complete KGs that instantiate all identified COPs.

End-to-End Pipeline Testing: Comprehensive testing over the Pilot Corpus processes documents through the complete integrated pipeline, revealing systematic issues including interaction effects between COP extractors, inconsistent tool coverage across discourse types, and representation generation errors. Testing systematically evaluates performance across document types and semantic phenomena, with particular attention to error propagation through pipeline stages.

From MWA to Production Annotation Model: Based on integration testing results, the MWA evolves into a production-ready annotation model suitable for creating comprehensive ground truth. This refinement may involve adding elements crucial for full corpus annotation such as coreference chains spanning multiple COPs, disambiguation tags, or confidence indicators, while maintaining backward compatibility with the ontology. The production annotation model serves as the schema for ground truth creation in Task 5.3.5.

Mapping Algorithm Enhancement: The RDF mapping procedures validated for individual COPs in Task 5.3.2 are consolidated and enhanced to handle the complete KG structure. This includes improving handling of complex semantic structures that emerge from COP interactions, adding validation using tools such as SHACL, OWL reasoners, SPARQLAnything [12], and RML [50], and implementing error handling mechanisms that manage extraction failures and partial results.

Following successful integration and refinement, the methodology proceeds to full corpus processing with comprehensive ground truth creation and systematic evaluation.

5.3.5 Knowledge Extraction and Evaluation (Task V)

Task V represents the transition from pilot development to full corpus processing. The refined system from Task 5.3.4 is applied to test data separate from

the Pilot Corpus, employing technical metrics and domain expert evaluation to assess whether extracted KGs accurately represent source discourse.

Ground Truth Preparation: Comprehensive ground truth is created by applying the production annotation model to a test dataset separate from the Pilot Corpus. This annotation follows the same paradigms established in Task 5.3.2, including multiple annotators and inter-annotator agreement measurement when resources permit. The ground truth encompasses all COPs and serves as the gold standard for systematic evaluation. The size of the test dataset should balance evaluation rigor with annotation resource constraints, typically ranging from 10-50 documents depending on document length and complexity.

Knowledge Extraction: Test datasets are processed through the complete integrated pipeline under realistic deployment conditions, with systematic documentation of performance and failure modes. This represents the first application of the extraction pipeline beyond the Pilot Corpus used for development.

Multi-Level Evaluation: Multiple complementary approaches address KG evaluation challenges:

- **Technical Evaluation**: Component-level assessment using precision, recall, and F_1-score evaluates individual extraction tasks independently for each COP. Coverage analysis examines whether the KG contains sufficient information to answer the original CQs;
- **Semantic Evaluation**: KG rehydration [73, 70] as described above;
- **Competency-Based Evaluation**: SPARQL query suites derived from original CQs verify that the KG satisfies the functional requirements that motivated its construction, aligned with [158] evaluation using tools like TestaLOD [32].

Domain Expert Validation: Comprehensive review by domain specialists evaluates extraction quality and coherence. The rehydration technique

enables evaluation by experts without RDF expertise by presenting KG content as natural language for assessment.

Iteration Strategy: Evaluation results may trigger returns to earlier tasks: coverage issues may necessitate returning to Task 5.3.2 for annotation schema refinement or to Task 5.3.4 for integration adjustments; extraction bottlenecks may require revisiting Task 5.3.3 for pipeline architecture modifications; systematic errors revealing ontological misalignments may necessitate returning to Task 5.3.1 for COP reassessment.

5.4 Conclusion

This chapter introduced ATR4CH, a systematic methodology designed to address the methodological challenges of text-to-KG development where specialized ontologies and limited training data intersect. The methodology emerged from recognizing that KE constitutes an active component of the digital curation life cycle, one that must be designed with explicit awareness of how technical decisions shape what knowledge becomes computationally accessible. ATR4CH positions KE within the curatorial framework established by [92, 194, 201], treating pipeline design as interpretative practice rather than purely technical implementation. The methodology addresses the bidirectional tension between ontological representation and technical extractability by taking ontological models as foundational inputs and systematically adapting annotation schemas and extraction architectures to align with predetermined representational frameworks. Even if the methodology was developed mainly for and inspired by the CH domain, it can be generalized to other use cases as long as they meet the input requirements. The methodology adapts principles from XD [158] to text-to-KG development through five iterative tasks. Task 5.3.1 establishes foundational understanding through parallel corpus and ontology analysis, identifying COPs that bridge CQs, ontological structures, and extractable textual content. Task 5.3.2 develops annotation schemas incrementally, processing COPs individually rather than attempting simultane-

ous coverage of entire ontologies. This pattern-by-pattern approach, validated through RDF mapping on the Pilot Corpus, produces the MWA that serves as the target schema for automated extraction. Task 5.3.3 implements extraction pipelines incrementally, developing and validating components for individual COPs before integration. Task 5.3.4 consolidates modular extractors into unified pipelines, refining the MWA into production annotation models suitable for comprehensive ground truth creation. Task 5.3.5 transitions from pilot development to full corpus processing through multi-level evaluation encompassing component-level metrics, semantic fidelity assessment through rehydration, CQ-based validation, and domain expert review. ATR4CH explicitly incorporates LLMs within text-to-KG pipelines to address data scarcity in underrepresented domains through few-shot and zero-shot learning [27]. The methodology positions annotation as serving dual purposes: as ground truth for evaluating models and as training data enabling supervised fine-tuning when computational resources permit. This dual function transforms annotation from mere data preparation into intensive interpretative engagement that develops domain expertise while revealing extraction challenges and potential ontological misalignments.

Chapter 6 implements the methodology for opinion mining in authenticity assessment debates, developing annotation models and extraction pipelines for complex scholarly discourse that exemplifies the multi-perspectival, evidence-based reasoning characteristic of CH interpretative scholarship.

Chapter 6

Ontology-driven Opinion Mining in Authenticity Assessment Debates

This chapter presents the implementation of ATR4CH (presented in Chapter 5) to extract scholarly opinions from authenticity assessment debates, validating the methodology through a domain that exemplifies complex evidential reasoning, multi-perspectival structures, and competing hypotheses characteristic of Cultural Heritage (CH) interpretative scholarship. The development of this project starts from annotation using INCEpTION[1] to a text-to-KG (Knowledge Graph) pipeline based on LLMs.[2]

Authenticity assessment debates constitute a fundamental aspect of CH scholarship, where scholars from different humanities disciplines (e.g., Diplomatics, Palaeography, Philology, History) and scientific fields (e.g., Forensics, Materials science, Chemical analysis) examine objects to determine their genuineness. Historical uncertainty, gaps in documentary transmission, and in-

[1] `https://inception-project.github.io/`

[2] This chapter draws from prior work by the author: Schimmenti, A., Pasqual, V., Tomasi, F., Vitali, F., & van Erp, M., *Structuring Authenticity Assessments on Historical Documents using LLMs*, in *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024* (2024), pp. 463–468. This preliminary work was significantly expanded in Schimmenti, A., Pasqual, V., Vitali, F., & van Erp, M., *Knowledge Graphs Generation from Cultural Heritage Texts: Combining LLMs and Ontological Engineering for Scholarly Debates* (2025, accepted in *Journal of Documentation*, DOI:`https://doi.org/10.1108/JD-07-2025-0203`)

terpretative subjectivity contribute to the inherent complexity of these debates [18, 24, 68]. Scholars frequently arrive at divergent conclusions based on different evidential priorities: paleographers examine script characteristics, philologists analyze linguistic patterns, historians contextualize claims within broader political narratives, and materials scientists assess physical properties. These debates embed rich knowledge structures within natural language discourse: multiple scholarly agents express competing assessments, each supported by distinct evidence drawn from intrinsic features (content, language, style), extrinsic features (materials, physical characteristics), and provenance information (historical context, transmission history).

Consider the *Donation of Constantine*, a purported 4th-century decree by Emperor Constantine transferring authority over Rome and the western Roman Empire to the Pope. In the 15th century, Lorenzo Valla exposed the document as a forgery through philological analysis, demonstrating that its Latin contained anachronisms from the 8th rather than 4th century [204]. Despite Valla's compelling evidence, acceptance of this finding evolved gradually over centuries, involving competing scholarly interpretations about the document's actual date of creation, its true author, and the intentions behind its fabrication. Different scholars proposed alternative hypotheses: some argued for 8th-century Papal authorship to legitimize territorial claims, others suggested Frankish origin to support Carolingian political interests, while some maintained modified authenticity claims arguing the text preserved genuine 4th-century material despite later interpolations [180].

Despite the rich scholarly discourse surrounding authenticity debates, existing knowledge bases fail to capture this interpretative complexity. As shown in Figure 6.1, Wikidata categorizes the *Donation of Constantine* as a "historical forgery"[3] with no representation of the scholarly debate surrounding this assessment. DBpedia[4] similarly lacks a structured representation of the authenticity discourse. In contrast, the corresponding Wikipedia page contains

[3]Donation of Constantine - Q238476
[4]Donation of Constantine - DBpedia entry

extensive discussions of Valla's philological arguments, the specific linguistic evidence he identified (anachronistic terminology, stylistic inconsistencies with 4th-century Latin), the institutional resistance from the Church, subsequent scholarly confirmations through independent analyses, and competing hypotheses about the forgery's actual date and authorship.[5]

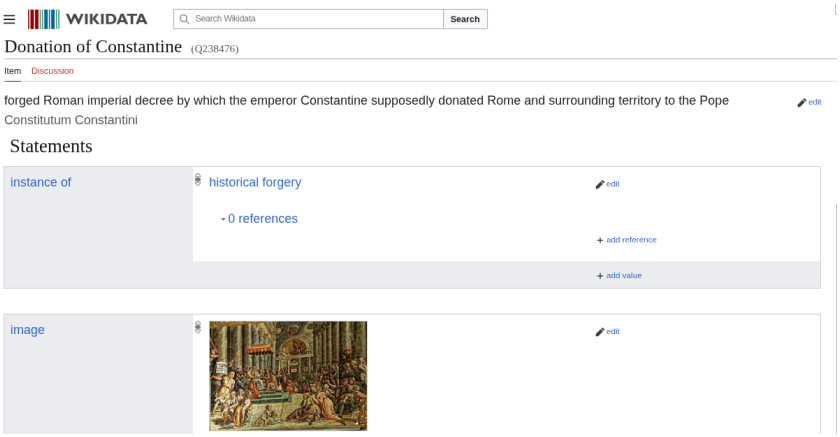


Figure 6.1: The Donation of Constantine entry in Wikidata

This misalignment between rich textual content and sparse structured claims is systematic across authenticity debates. Whether examining literary pseudepigrapha, archaeological forgeries, or art attribution controversies, complex scholarly reasoning gets reduced to simple categorical assertions that fail to capture the evidential reasoning, methodological disagreements, and evolving consensus that characterize authentic scholarly discourse. The challenge is twofold: first, representing competing scholarly opinions within formal knowledge representation systems requires sophisticated mechanisms that traditional implementations struggle to handle effectively; second, extracting this complex scholarly information from textual sources at scale requires enormous manual labor, creating insurmountable scalability barriers for most CH institutions [147, 180].

[5]The Donation of Constantine page on Wikipedia. Last visited: October 2025

6.1 Problem Statement

This case study addresses the challenge of extracting and structuring multi-perspectival scholarly interpretations from authenticity assessment debates. The extraction targets four interconnected knowledge layers:

1. **CH Item Metadata**—alleged information the object claims about itself (creator, date, location) before scholarly critical analysis;

2. **Scholarly Opinions**—authenticity assessments expressed by scholarly agents, classified as `Authentic`, `Forgery`, `FormalForgery`, `ContentForgery`;

3. **Evidential Features**—specific characteristics examined by scholars to support assessments, organized by type with evaluations;

4. **Alternative Hypotheses**—competing scholarly claims about the object's actual creator, date, location, or intended purpose.

These layers are represented using the SEBI (Scholarly Evidence-Based Interpretation)[147] ontology[6], which employs RDF-star reification to capture multiple concurrent claims with their contextual information.

The corpus comprises 581 Wikipedia articles describing historical forgeries, hoaxes, and authenticity debates across CH domains, including literary forgeries (138 articles), pseudepigraphy (65), archaeological forgeries (52), document forgeries (33), and art-related controversies. These articles average 8,150 characters and exhibit substantial lexical diversity, with discourse patterns ranging from straightforward scholarly consensus to complex multi-party debates spanning centuries.

6.2 Research Questions

This implementation addresses the following specific questions derived from the primary research questions established in the Introduction:

[6] `https://valentinapasqual.github.io/sebi/`

- **RQ1: Extraction Performance**: How accurately can systematic LLM-based pipelines extract different components of scholarly discourse in authenticity debates, including metadata, agents, evidential reasoning, and interpretative hypotheses?

- **RQ2: Representation Fidelity**: Do automatically generated KGs adequately represent the complexity and nuance of authenticity assessment interpretations when following the ATR4CH approach?

- **RQ3: Model Comparison**: How do different LLMs perform within structured extraction pipelines for authenticity assessment texts, and what are the implications for cost-effective deployment in CH institutions?

- **RQ4: Methodology Validation**: What insights does this case study provide about ATR4CH's broader applicability to other forms of CH interpretative scholarship?

6.3 Outline

Section 6.4 reviews related work on knowledge representation for interpretative scholarship and opinion mining in CH contexts. Section 6.5 describes the construction of the corpus, the analysis of the SEBI ontology, and the development of the annotation model. Section 6.6 presents the LLM-based text-to-KG pipeline. Section 6.7 provides comprehensive evaluation results across five dimensions, comparing performance of Claude Sonnet 3.7, Llama 3.3 70B, and GPT-4o-mini. Section 6.8 synthesizes findings in relation to the research questions, analyzing performance trade-offs, deployment implications, and contributions. The code of the annotation process, text-to-KG pipeline and evaluation is available as a GitHub repository.[7]

[7] `https://github.com/aschimmenti/SEBI-Knowledge-Extraction.git`

6.4 Related Work

The challenges of representing and extracting interpretative knowledge in the CH domain have received increasing attention in recent research, particularly concerning Knowledge Representation (KR) and Knowledge Extraction (KE) tasks. We focus first on conceptual and ontological models developed for multi-perspective KR, and then turn to methods for extracting such interpretations from unstructured texts, including recent advances in LLMs.

6.4.1 Knowledge Representation

Recent theoretical advancements address the complex epistemic characteristics inherent in CH scholarship, particularly distinguishing between uncertainty and interpretative multiplicity [154]. These concepts, while related, operate at different epistemological levels. Uncertainty concerns the degree of confidence in factual claims—for instance, paleographic analysis might establish that a manuscript dates to 1450–1470 CE with 80% confidence, reflecting incomplete evidence about an objective historical fact. In contrast, interpretative knowledge represents scholarly positions grounded in evidential reasoning where different experts, examining the same evidence through distinct methodological frameworks, arrive at competing but equally defensible conclusions. When two art historians disagree about whether a painting's composition reflects Caravaggio's influence, both interpretations may be well-supported by formal analysis, historical context, and comparative evidence; the multiplicity arises not from uncertainty about facts but from legitimate differences in interpretative frameworks and evidential weighting [153]. The epistemic challenge for knowledge representation lies in distinguishing these layers: capturing that "Scholar A interprets feature X as evidence for hypothesis Y" constitutes certain knowledge about the interpretation itself, even when the truth value of hypothesis Y remains contested. A KG encoding authenticity debates must represent with full certainty that Lorenzo Valla argued the Donation of Constantine was an 8th-century forgery based on linguistic anachronisms, while simultaneously acknowledging that competing scholars proposed alternative

dating and authorship hypotheses. The assertional content ("the document is an 8th-century forgery") carries uncertainty regarding historical truth, but the meta-level claim ("Valla interpreted the evidence as indicating 8th-century forgery") represents certain knowledge about scholarly discourse. However, these theoretical advances have not been translated into widely adopted practical tools and standards in KGs. Standard online catalogs (e.g. Europeana)[8] typically provide flat metadata with a single-perspective, relegating discussions, debates, and uncertain facts to free text descriptions [16].

To the best of our knowledge, Wikidata is the only large-scale data catalog that employs a custom reification method to integrate claims with varying degrees of truthfulness, i.e. its ranking mechanism. Despite the adequate expressive power made available by the Wikidata model, annotators in the CH domain under-use this feature. Additionally, claims related to CH data often make use of numerous qualifiers to encode contextual metadata, likely due to the greater effort required for this type of annotation [47].

Different ontologies have been designed to structure multi-perspective representations in CH data. ICON [173, 17] encodes visual recognitions in art history using *n*-ary relations to encode contextual metadata. Digital Hermeneutics [40] employs a layered approach using Named Graphs [33] to represent scholarly interpretations in archival and literary sources. HiCo [41] and the STAR model [11] have been designed to represent historical interpretations and arguments. Wider adoption is hampered by the lack of tools to support the extraction, categorization, and contextualization of scholarly interpretations on a scale, where practical effectiveness ultimately depends on the ability to extract such interpretations from unstructured sources. This is the gap that the work presented in this paper aims to fill: it introduces a method for populating an ontology of interpretative claims, addressing the need for representations of scholarly discourse within CH datasets.

[8] `https://www.europeana.eu/`

6.4.2 Opinion Mining for the Semantic Web

Opinion mining is a task which has seen rare application in CH contexts. Aspect-Based Sentiment Analysis (ABSA) through SemEval 2014 [156] established granular opinion extraction frameworks with four subtasks: aspect term extraction, aspect term polarity, aspect category detection, and aspect category polarity. Datasets for this task are mainly for product and restaurants reviews [178]. Recent advancements expanded the possible approaches [52, 72, 86], with context-aware language models such as BERT improving performance [72] and LLMs such as GPT-3.5 achieving state-of-the-art results with zero-shot prompting [212]. As for open text-to-KG approaches, Sentilo [162] is a system tailored for sentiment analysis based on a similar architecture as text2AMR2FRED [71]. Overall, an ABSA-oriented approach could be transferred to scholarly opinions within a debate, as analyzing the sentiment and the specific aspects of a review is not structurally dissimilar to a scholarly opinion, where some evidences (parallel to aspects) are followed by an assertion (parallel to sentiment).

6.5 Methodology and Materials

This section describes the implementation of ATR4CH (Chapter 5) for authenticity assessment debates. Following the five-task structure, we first establish the foundational inputs (Section 6.5.1), then present the iterative development of annotation schemas (Section 6.5.2), and finally describe the consolidated production annotation model (Section 6.5.3).

6.5.1 Foundational Inputs (Task I)

ATR4CH requires three foundational inputs: a corpus of unstructured documents, a target ontology defining knowledge representation (SEBI), and Competency Questions (CQs) specifying information requirements. This subsection presents these inputs and the analytical activities of Task I.

Corpus Collection and Analysis

The corpus comprises Wikipedia articles focusing on historical forgeries, hoaxes, and authenticity controversies across CH domains. The dataset was collected using web scraping from Wikipedia's categorical organization system, targeting categories related to forgeries and authenticity debates. Articles were retrieved from 15 categories on Wikipedia, with full article text and internal links (sitelinks[9]). The script selected both categorical pages[10] and standalone articles,[11] storing each document with complete textual content and associated metadata including categorization and cross-references to related entities. The initial selection covered 31 categories, including Document[12] and Literary Forgeries,[13] Historical Myths,[14] Conspiracy Theories,[15] Pseudepigraphy (texts whose claimed author differs from the actual author, or works whose real author attributed them to historical figures),[16] and Political forgery.[17] From the retrieved 1301 documents,[18] 16 categories and 717 articles were excluded because they presented no scholarly debate. The final dataset encompasses 581 articles as shown in Table 6.1.

Corpus Characteristics. The corpus exhibits variability in document length and complexity, with articles averaging 8,150 characters and 1,249 tokens per document. Unique vocabulary per article averages 464 tokens, indicating substantial lexical diversity within authenticity assessment discourse. As shown in Figure 6.2, the distribution of the length of the articles follows a distorted pattern on the right, with most articles ranging from 2k to 15k characters and outliers extending beyond 40k characters.

[9] `https://www.wikidata.org/wiki/Help:Sitelinks`

[10] For instance, the Wikipedia Category "Forgery" (`https://en.wikipedia.org/wiki/Category:Forgery`)

[11] For instance, the article describing the Donation of Constantine (`https://en.wikipedia.org/wiki/Donation_of_Constantine`)

[12] `https://en.wikipedia.org/wiki/Category:Document_forgeries`

[13] `https://en.wikipedia.org/wiki/Category:Literary_forgeries`

[14] `https://en.wikipedia.org/wiki/Category:Historical_myths`

[15] `https://en.wikipedia.org/wiki/Category:Conspiracy_theories`

[16] `https://en.wikipedia.org/wiki/Category:Pseudepigraphy`

[17] `https://en.wikipedia.org/wiki/Category:Political_forgery`

[18] Selection performed October 2024

Table 6.1: Distribution of articles across Wikipedia categories in the corpus

Category	Article Count
Literary forgeries	138
Pseudepigraphy	65
Old Testament pseudepigrapha	60
Forgery controversies	58
Archaeological forgeries	52
Musical hoaxes	44
Art forgers	40
Document forgeries	33
Ancient Greek pseudepigrapha	28
Political forgery	26
Religious hoaxes	15
Modern pseudepigrapha	11
Sculpture forgeries	7
Political forgeries	2
Shakespeare authorship question	2
Total	**581**

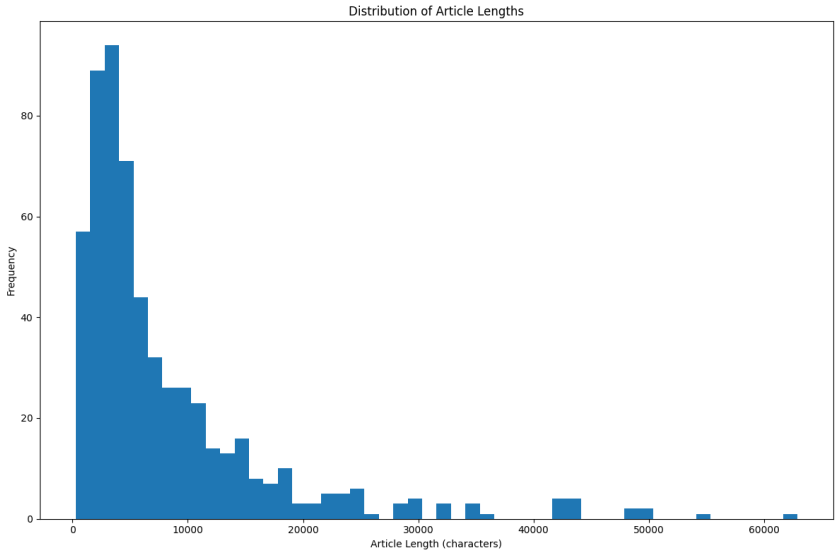


Figure 6.2: Overall distribution of article lengths showing the right-skewed pattern characteristic of encyclopedic content, with most articles in the 2k–15k character range and notable outliers extending beyond 40k characters

The distribution among categories reflects the natural prevalence of different types of forgery in the academic discourse (Figure 6.3). Literary forgeries represent the largest category with 138 articles, followed by pseudepigraphy

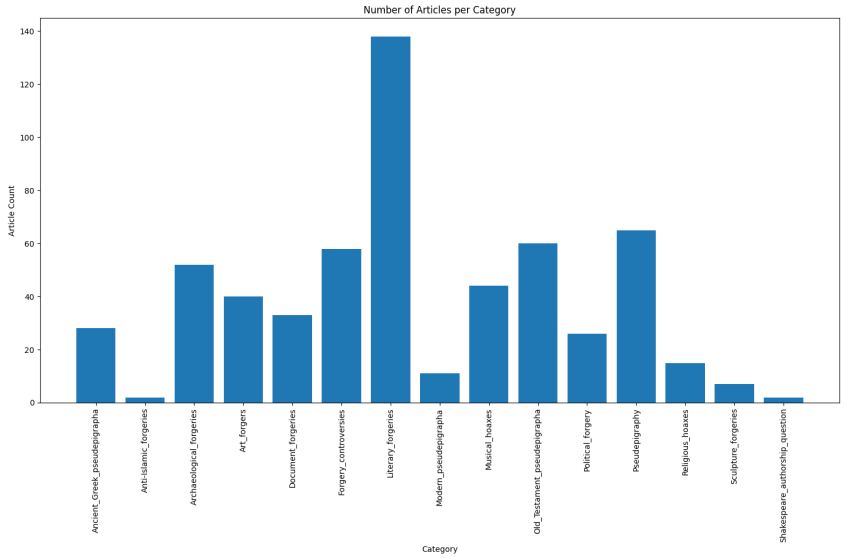


Figure 6.3: Distribution of articles across Wikipedia categories, showing the natural prevalence of different forgery types in scholarly discourse

forms totaling 136 articles across subcategories. Archaeological and artistic forgeries comprise 132 articles combined, while specialized categories such as musical hoaxes and religious controversies contain fewer but often more detailed entries.

The temporal scope spans from late antiquity to the contemporary period. Token count distribution by category (Figure 6.4) reveals substantial variability, with numerous outliers indicating comprehensive case studies. The Shakespeare authorship question category demonstrates the highest token density, with articles reaching nearly 10K tokens. Political forgery and religious hoaxes also show elevated token counts. Categories like musical hoaxes and modern pseudepigrapha exhibit more consistent, moderate-length articles with fewer outliers.

Notable corpus examples include the *Demodocus*[19], a fabricated Platonic dialogue exemplifying early pseudepigraphic practices. This is a counterexample: the relative Wikidata entry exists[20] correctly employs a deprecated rank for the authorship claim linking Plato to the Demodocus (Figure 6.5).

[19] `https://en.wikipedia.org/wiki/Demodocus_(dialogue)`
[20] `https://www.wikidata.org/wiki/Q2625856`

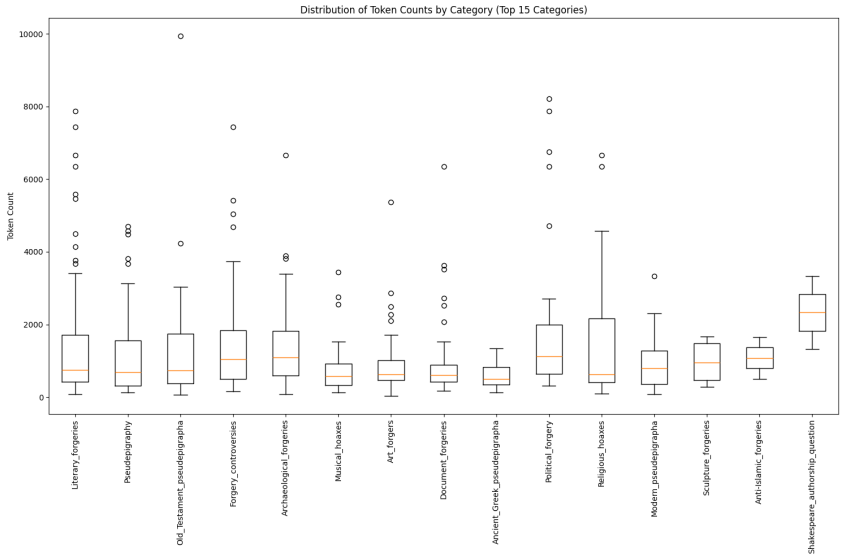


Figure 6.4: Token count distribution by category, illustrating variability in article length and content density. Box plots show medians, quartiles, and outliers representing comprehensive case studies

Figure 6.5: Plato noted as the author of the Demodocus using a deprecated rank, illustrating how existing knowledge bases can represent disputed attributions

The corpus also includes the *Protocols of the Elders of Zion*, one of the most notorious forgeries[21]. Among the most recent examples, the 1996 *Posthumous Diary*[22], allegedly forged poems by Italian poet Eugenio Montale that generated debate in the Italian philology community.

Discourse Analysis. Following Task I guidelines (Section 5.3.1), corpus analysis identified linguistic patterns and discourse structures that manifest authenticity assessment knowledge. Key challenges include implicit mentions requiring contextual inference, long-distance dependencies where KG components are separated by substantial text spans, nested entities discussed through relational structures, and ambiguous references using informal names or fig-

[21] `https://en.wikipedia.org/wiki/The_Protocols_of_the_Elders_of_Zion`
[22] `https://en.wikipedia.org/wiki/Posthumous_Diary`

ures of speech. In Wikipedia articles about forged CH items, this qualitative analysis revealed patterns in both structure and content: sections containing scholarly opinions versus out-of-scope debates, enabling focused extraction on relevant passages.

Ontology Analysis

The Scholarly Evidence Based Interpretation ontology (SEBI)[23] [147] was developed to represent opinions from scholarly articles (e.g., [89]), a catalogue describing 153 known forgeries from Styria [85], and discussions with an expert diplomatist [147]. The data model represents authenticity assessment claims using RDF-star [90] as a reification method to represent (possibly concurrent) claim contents and contextual information [147].

Each claim provides information about the document: authenticity classification, date and place of creation, author, and intention behind creation. Contextual information about the claim includes evidence collected by the scholar to reach conclusions using evidence-based evaluations and the author of the claim with relevant bibliographic entries (using HiCo [24] and PROV-O[25]. RDF-star [90] was chosen as the reification method to express both the content of the claim and the context, allowing the representation of the complete evaluation process conducted by the scholars.

As shown in Figure 6.6, each claim contains the pattern to classify the authenticity. Items can be instances of one of the classes `sebi:Forgery`, `sebi:Authentic`, `sebi:FormalForgery`, `sebi:ContentForgery`, all subclasses of `sebi:Document`.

Additionally, each RDF-star quoted triple includes details such as the believed creator of the document (expressed through the triple `sebi:Document`–`dct:creator`–`dct:Agent`), date of creation (`sebi:Document`–`dct:date`–`time:Interval`), location of creation (`sebi:Document`–`dct:coverage`–`dct:Location`), and intention behind document creation (`sebi:Document`–

[23] `https://valentinapasqual.github.io/sebi/`
[24] `https://marilenadaquino.github.io/hico/`
[25] `https://www.w3.org/TR/prov-o/`

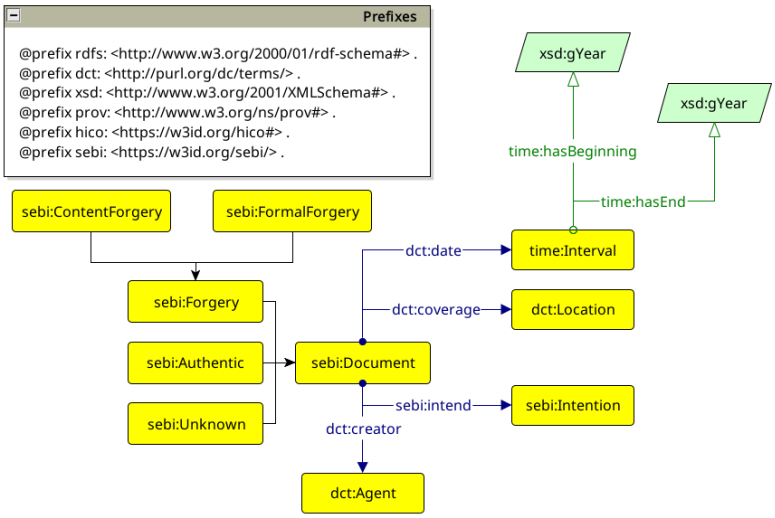


Figure 6.6: Selection of classes and properties to represent scholarly claims addressing authenticity assessment of a document

`sebi:intended`–`sebi:Intention`). The `dct:date` property connects to a `time:Interval` class, which includes `time:hasBeginning` and `time:hasEnd` properties to specify creation periods and handle fuzzy time-spans [147].

Concerning contextual information (Figure 6.7), each interpretation (set of claims represented as quoted triples) is categorized as a `hico:InterpretationAct` connected to a `prov:Agent` to address its authoriality and linked to evidence supporting the claim (`sebi:support sebi:Evidence`) [147].

Document features and their evaluation are components of the ontology. Document features (`sebi:Feature`) are either extrinsic features (`sebi:ExtrinsicFeature`), intrinsic ones (`sebi:IntrinsicFeature`), or provenance information (`sebi:Provenance`), capturing aspects such as ink, support, handwriting, and orthography. Each feature is evaluated on established criteria (`sebi:Evidence`) such as consistency, presence, completeness, veridicality, and reliability. A score is associated with each evidence as `xsd:Literal` using the property `forgont:hasEvaluationScore`. The evaluation score indicates a measure on each collected evidence, allowing integration of negatives (e.g., the absence of signature in a document is represented as ev-

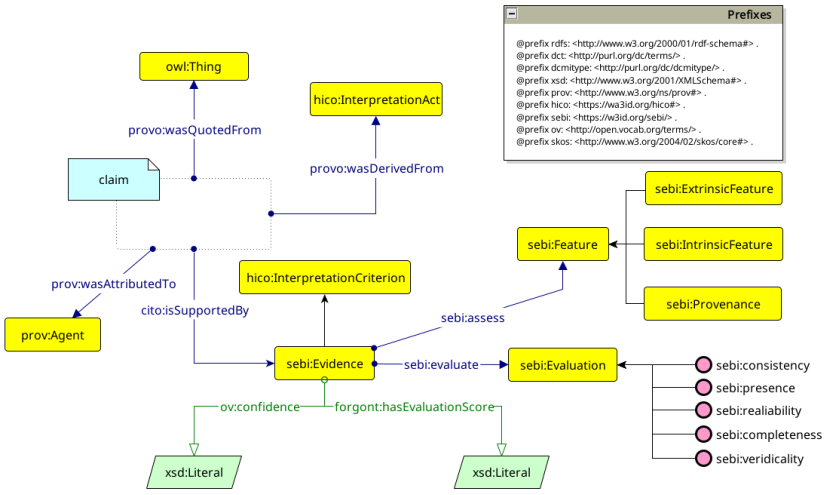


Figure 6.7: Selection of classes and properties to represent contextual information about scholarly claims addressing authenticity assessment of a document

idence based on the feature "authentication marks" with evaluation "presence", with score `false` or 0) [147].

Core Ontological Patterns Identification

Following Task I, four Core Ontological Patterns (COPs) were identified for extraction, presented here in hierarchical priority order:

1. **CH Item Metadata**—Alleged information that the object "claims" about itself (creator, date, location) before scholarly critical analysis

2. **Scholarly Opinions**—Authenticity assessments expressed by scholarly agents, classified as `Authentic`, `Forgery`, `FormalForgery`, `ContentForgery`

3. **Evidential Features**—Characteristics examined by scholars to support assessments, organized by type with evaluations

4. **Alternative Hypotheses**—Competing scholarly claims about the object's actual creator, date, location, or intended purpose

These COPs emerged from intersection of CQs, ontological structures, and extractable content patterns identified in corpus analysis.

Pilot Corpus Selection

Following Task I guidelines, a Pilot Corpus was selected for iterative development. Seven articles were chosen (*Donation of Constantine*,[26] *Eremin Letter*,[27] *Getty Kouros*,[28] *Historia Augusta*,[29] *Life of Homer*,[30] *Marriage Charter of Empress Theophanu*,[31] *Protocols of the Elders of Sion*[32]), each belonging to a different category. These articles were selected based on the following criteria: (1) presence of multiple scholarly perspectives on authenticity, (2) clear attribution of claims to specific researchers or institutions, (3) discussion of evidence-based reasoning, and (4) representation of different temporal periods and document types.

6.5.2 Minimal Working Annotation Development (Task II)

Following the pattern-by-pattern approach specified in Task II (Section 5.3.2), annotation schemas were developed iteratively for each COP. This subsection presents the initial Minimal Working Annotation (MWA) focusing on the highest-priority COP: CH Item Metadata and Scholarly Opinions.

Initial Annotation Schema

The annotation model was developed through INCEpTION [109], implementing three core patterns from SEBI: **CH item metadata**, **scholarly agents**, and **authenticity opinions**.

CH Item Metadata. An identification layer was established using INCEpTION's Knowledge Base integration with Wikidata, allowing annotators to link textual mentions directly to Wikidata IDs for automatic coreference resolution. Item types mentioned in source texts were reconciled to DCMI Type Vocabulary classes (`dcmitype:Text`, `dcmitype:PhysicalObject`,

[26] `https://en.wikipedia.org/wiki/Donation_of_Constantine`
[27] `https://en.wikipedia.org/wiki/Eremin_letter`
[28] `https://en.wikipedia.org/wiki/Getty_kouros`
[29] `https://en.wikipedia.org/wiki/Historia_Augusta`
[30] `https://en.wikipedia.org/wiki/Life_of_Homer_(Pseudo-Herodotus)`
[31] `https://en.wikipedia.org/wiki/Marriage_Charter_of_Empress_Theophanu`
[32] `https://en.wikipedia.org/wiki/The_Protocols_of_the_Elders_of_Zion`

`dcmitype:Collection`) with appropriate subclass relationships.

Scholarly Agents. Entities expressing opinions (Cognizers) correspond to `dct:Agent` in the ontology. Each Cognizer was linked to Wikidata when possible, with fallback strategies for entities without entries, impersonal statements, and consensus attributions.

Authenticity Claims. Claims were modeled through directed relations between Cognizer spans and CH item spans, labeled according to SEBI's authenticity categories (`Authentic`, `FormalForgery`, `ContentForgery`, `Forgery`, `Neutral`). Each opinion becomes an RDF-star quoted triple linked to `hico:InterpretationAct`.

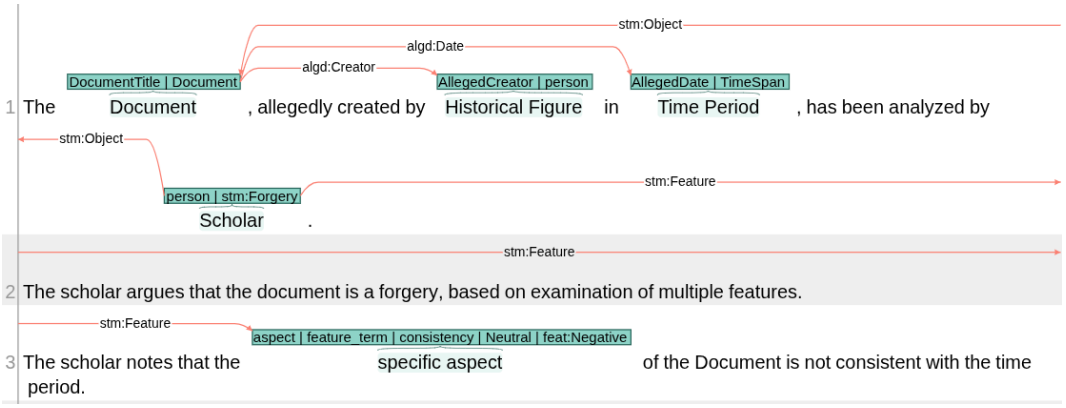


Figure 6.8: Example annotation of an entity expressing an opinion about a CH item

RDF Mapping Validation

Using this approach on the Pilot Corpus, annotation-to-RDF mapping was validated through the algorithm in Listing 6.1. This mapping served as a unit test for the annotation schema, validating that individual annotation patterns correctly transformed to valid RDF instantiating SEBI classes and properties.

Listing 6.1: Core Annotation Mapping Algorithm

```
STEP 1: Extract Cognizer-Opinion Pairs
Select all spans marked as Entity
WHERE span also has Opinion tagset label
=> CognizerSet(Cognizer(CognizerSpan, Opinion, WikidataID)
```

```
5
6  STEP 2: Extract CH Items
7  Select all spans marked as Entity
8  WHERE span has ItemTitle label
9  => ItemSet(ItemSpan, WikidataID)
10
11 STEP 3: Find Relations
12 For CognizerSpan in CognizerSet, check if CognizerSpan
13 has stm:Object relation to span in ItemSet
14 => Valid tuples (Cognizer, Item, Opinion)
15
16 STEP 4: Generate RDF for each tuple
17 For each matching pattern:
18 |-- Generate URI for Cognizer
19 |-- Add owl:sameAs + Wikidata ID
20 |-- Generate URI for Item
21 |-- Add owl:sameAs + Wikidata ID
22 |-- Map opinion to corresponding SEBI class (e.g., sebi:Forgery)
23 |-- Generate URI for Named Graph (hico:InterpretationAct)
24 |-- Generate claim triple as a RDF-star statement
25
26 +-- Apply template:
27
28     ex:{cognizer_uri}_about_{item_uri} rdf:type hico:InterpretationAct
       ;
29     prov:wasAttributedTo ex:cognizer .
30
31     ex:cognizer rdf:type dct:Agent ;
32     rdfs:label "CognizerSpan"@language ;
33     owl:sameAs wd:wikidataId .
34
35     ex:item rdf:type ex:type ;
```

```
36      rdfs:label "ItemSpan"@language ;
37      owl:sameAs wd:wikidataID .
38
39      << ex:item rdf:type sebi:Opinion >> prov:wasDerivedFrom ex:{
        cognizer_uri}_about_{item_uri} .
```

Successful RDF generation from annotations validated the initial schema design, confirming that the annotation patterns adequately captured the semantic content required to instantiate SEBI's authenticity claim structure.

6.5.3 Production Annotation Model (Task IV)

Following successful development and validation of extraction pipelines for individual COPs (Tasks II–III, detailed in Section 6.6.1), the annotation schemas were consolidated into a production model suitable for comprehensive ground truth creation. This refined model captures CH item metadata, evidence and features, and scholarly hypotheses through additional layers developed iteratively during pipeline testing on the Pilot Corpus.

CH Item Metadata Layer

This layer captures *alleged metadata*—descriptive information (creator, date, location) that the document or artifact purports about itself, before scholarly critical analysis. This includes face-value claims presented within the item or by who claimed to find the item regarding authorship, creation date, geographic origin, and other identifying characteristics. Annotations include `AllegedCreator`, `AllegedDate`, `AllegedLocation`, `ItemSubject`, and `ItemType`, plus properties for formal forgeries (`ItemCreator`, `ItemDate`, `ItemLocation`).

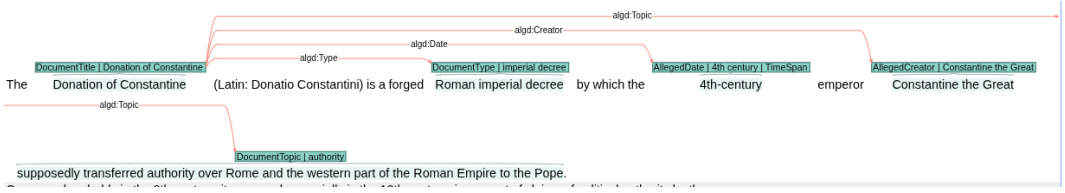


Figure 6.9: Alleged metadata annotation for the Donation of Constantine

Evidence and Features Layer

This layer generates Evidence nodes connected to InterpretationAct Named Graphs. It employs four tagsets: Feature (SEBI vocabulary terms for intrinsic/extrinsic features and provenance), FeatureAssessment (evaluation perspectives: consistency, presence, completeness, reliability, veridicality), FeatureAssessmentPolarity (negative, neutral, positive), and FeatureAssessmentConfidence.

Consider Lorenzo Valla's assessment of the Donation's language features (Figure 6.10), which converts to three evidence structures linking textual features to evaluation criteria and polarities.

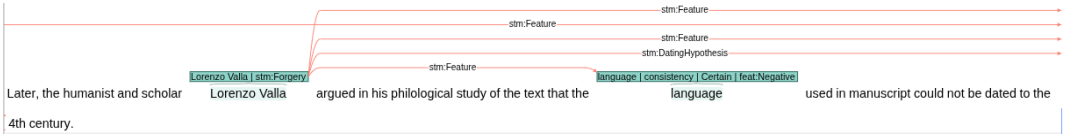


Figure 6.10: Lorenzo Valla's opinion with feature assessment annotation

Listing 6.2 shows the evidence mapping algorithm.

Listing 6.2: Evidence and Feature Mapping Algorithm

```
STEP 1: Extract Evaluated Features
Select all spans marked as feature
WHERE span also has FeatureAssessment label, FeatureAssessmentPolarity
    , FeatureAssessmentConfidence
=> FeatureSet(FeatureSpan, FeatureClass, FeatureAssessment,
    FeatureAssessmentPolarity, FeatureAssessmentConfidence)

STEP 2: Select all spans marked as Entity
WHERE span also has Opinion tagset label
=> CognizerSet(Cognizer(CognizerSpan, Opinion, WikidataID)

STEP 3: Find Relations
For FeatureSpan in FeatureSet, check if CognizerSpan
has stm:Feature relation to any span(s) in FeatureSet
```

```
=> Valid tuples (Cognizer, FeatureSet)

STEP 4: Generate nodes
For each matching pattern:
|-- Generate/Reuse URI for Cognizer
|--- Add owl:sameAs + Wikidata ID

|-- Generate URI for sebi:Evidence graph
|--- match FeatureAssessment individual with sebi:Evaluation
    individual
|--- match FeatureAssessmentPolarity
|--- attach FeatureAssessmentConfidence score

|-- Generate URI for sebi:Feature graph
|--- attach FeatureSpan through rdfs:label
|--- attach FeatureClass through skos:broader

STEP 5: Generate RDF graph

+-- Apply template:

kb:{cognizer_uri}_about_{item_uri}_{idx} a sebi:Evidence ;
    sebi:assess kb:{feature_uri} ;
    sebi:evaluate sebi:{evaluation_uri} ;
    sebi:hasEvaluationScore "{polarity}"@language ;
    sebi:support kb:interpretation_act ;
    ov:confidence 1.0 .

kb:{feature_uri} a sebi:Feature ;
    rdfs:label "{FeatureSpan}"@language ;
    sebi:isAssessedBy kb:{cognizer_uri}_about_{item_uri}_{idx} ;
    skos:broader sebi:{feature_vocabulary_term} .
```

Scholarly Hypotheses Layer

This layer captures alternative hypotheses through four relation types linking Cognizers to Wikidata entities: `stm:CreatorHypothesis`, `stm:DatingHypothesis`, `stm:LocationHypothesis`, and `stm:ReasonHypothesis`.

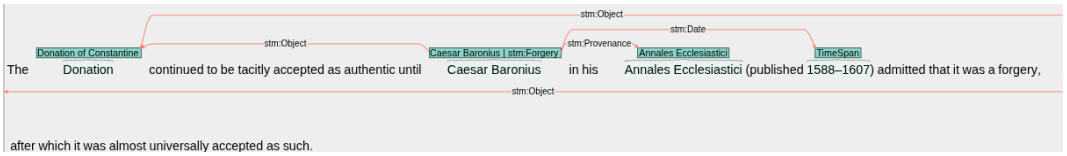


Figure 6.11: Caesar Baronius's admission of forgery with provenance annotation

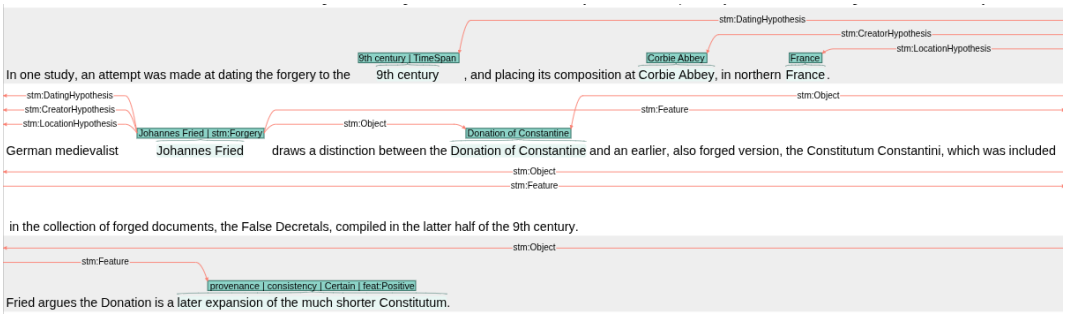


Figure 6.12: Johannes Fried's hypotheses annotation for the Donation of Constantine

Listing 6.3 details the hypotheses mapping algorithm.

Listing 6.3: Hypotheses Mapping Algorithm

```
STEP 1: Extract Hypothesis Relations
Select all relations of type:
|-- stm:CreatorHypothesis
|-- stm:DatingHypothesis
|-- stm:LocationHypothesis
|-- stm:ReasonHypothesis
=> HypothesesSet(CognizerSpan, HypothesisType, TargetSpan, WikidataID)

STEP 2: Extract Cognizer Entities
Select all spans marked as Entity
WHERE span also has Opinion tagset label
=> CognizerSet(CognizerSpan, Opinion, WikidataID)
```

```
13
14 STEP 3: Find Valid Patterns
15 For each relation in HypothesesSet:
16 Check if CognizerSpan exists in CognizerSet
17 => Valid tuples (Cognizer, HypothesisType, Target)
18
19 STEP 4: Generate Target URIs
20 For each matching pattern:
21 |-- Generate/Reuse URI for Cognizer
22 |-- Generate/Reuse URI for Item
23 |-- Generate/Reuse URI for Target entity
24 |-- Map HypothesisType to corresponding RDF property
25
26 STEP 5: Generate RDF-star Statements
27
28 +-- Apply template:
29
30 kb:{target_uri} a {target_class} ;
31     owl:sameAs wd:{wikidata_id} ;
32     # if Wikidata ID not available
33     # kb:{urifiedTargetSpan} a {target_uri} ;
34     rdfs:label "{TargetSpan}"@language .
35
36 << kb:{item_uri} dct:creator kb:{target_uri} >>
37     prov:wasDerivedFrom kb:{cognizer_uri}_about_{item_uri} .
38
39 << kb:{item_uri} dct:date kb:{target_uri} >>
40     prov:wasDerivedFrom kb:{cognizer_uri}_about_{item_uri} .
41
42 << kb:{item_uri} sebi:location kb:{target_uri} >>
43     prov:wasDerivedFrom kb:{cognizer_uri}_about_{item_uri} .
44
```

```
<< kb:{item_uri} sebi:intendedTo kb:{target_uri} >>
    prov:wasDerivedFrom kb:{cognizer_uri}_about_{item_uri} .
```

Ground Truth Statistics

Each annotation layer maps to RDF following the SEBI ontology principles, with Wikidata integration providing entity resolution. The INCEpTION project is available on GitHub alongside mapping scripts.[33] Statistics for the annotation results are shown in Table 6.2.

Table 6.2: Ground Truth annotation results

Span	Count
CH Items	45
Entities	235
Interpretation Acts	215
Evidences	132
Features	115
Wikidata alignments	308

6.6 Knowledge Extraction Pipeline and Evaluation Framework

This section presents the implementation of Steps III-V of the ATR4CH methodology (discussed in the previous chapter, Section 5.3.3 to 5.3.5), transforming the annotation model into a working KE pipeline. Our implementation integrates three complementary technologies in a sequential LLM-based process:

- GliNER for lightweight NER;
- LLMs for structured information extraction;
- Rule-based entity linking for external KG integration.

[33] SEBI-KE repository. See 'Inception2Graph' folder

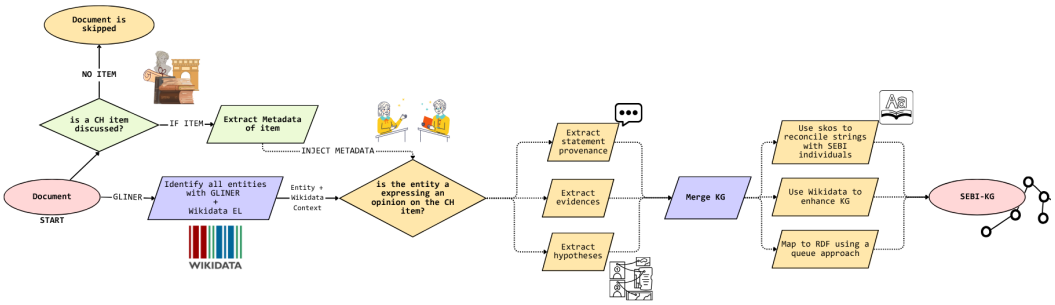


Figure 6.13: Flowchart of the sequential pipeline for SEBI-based KG generation

Each component addresses specific challenges in CH knowledge extraction while maintaining alignment with the SEBI ontology and supporting the complex semantic dependencies characteristic of humanities discourse.

GliNER [225] provides lightweight, generalist NER using custom entity types with state-of-the-art performance. **LLMs** handle structured information extraction through JSON schema-based responses.

The pipeline was used with three different models evaluated at varying parameter scales to understand performance trade-offs: Claude Sonnet 3.7[34], Llama 3.3 70B [56], and GPT-4o-mini.[35][36]. All models offer high performance in their respective size categories, reliable API availability, and built-in structured output support through tool calling interfaces.

Figure 6.13 presents a comprehensive overview of the sequential processing pipeline.

Entity Linking employs a rule-based approach leveraging the Wikibase API[37] and domain-specific heuristics. After testing various state-of-the-art solutions, this approach proved most effective for historical entities and CH concepts, providing reliable external knowledge base integration while handling the specialized vocabulary of authenticity assessment debates.

While the selected LLMs are capable of processing documents in their entirety at any step, the system automatically selects only the relevant paragraphs

[34] Claude Sonnet 3.7 Model Card
[35] GPT-4o-mini Model Card
[36] While the exact parameter size of GPT-4o-mini remains undisclosed, estimates range from 8-14 billion active parameters [19].
[37] Wikibase API Documentation

whenever possible. This design serves three strategic purposes: (1) reducing content volume per processing step to minimize potential opinion overlap between entities, assuming it improves precision; (2) demonstrating the pipeline's scalability to documents of arbitrary length; and (3) maintaining computational efficiency and cost-effectiveness by minimizing token consumption per API call.

6.6.1 Sequential Processing Pipeline

The KE pipeline consists of six components, each enriching the output before passing it to the next. Each component produces a JSON output following a predefined schema designed to be convertible to RDF. The development of the pipeline began with preliminary implementations and sequential testing over the Core Ontological Patterns (COPs) identified in Section 5.3.1 until the full KG could be extracted. As the output JSON model closely resembles the Ground Truth (GT) structure, the mapping uses similar logic, differing only in that it processes from JSON rather than the JSON UIMA CAS (Content Analysis System) format used by INCEpTION.

1. Raw text documents → metadata extraction → alleged/settled item metadata;

2. Item metadata + text → opinion holder identification → entity mentions with classifications;

3. Entity mentions → entity resolution → Wikidata-linked entity clusters;

4. Linked entities + paragraphs → opinion extraction → structured authenticity opinions;

5. Opinions + contexts → evidence mining → feature evaluations with polarity;

6. Evidence + full context → hypothesis extraction → conflicting statements.

CH Item Metadata Extraction

Aim: Identify all the CH items being discussed in an article and extract their alleged metadata

Input: Raw Wikipedia articles in markup (.txt files)

Output: Cleaned articles; JSON with alleged item metadata, based on a given JSON schema.

This component identifies and extracts metadata about CH items discussed in each article. The LLM is instructed to extract a JSON schema from the article text that describes all CH items under discussion. Specifically, it extracts the CH item *alleged metadata*—what items claim to be—including purported authors, creation dates, locations, item types, and subject matter. The task relies on In-Context Learning (ICL) in a Few-Shot setting (with 3 examples), using Chain-of-Thought (CoT) reasoning. Listing 6.14 shows an example text extracted from the Donation of Constantine Wikipedia article as input (on the left) and the JSON output (on the right).

Input source text

The **Donation of Constantine** [...] is a forged Roman imperial decree by which the 4th-century emperor Constantine the Great supposedly transferred authority over Rome and the western part of the Roman Empire to the Pope [...]. [I]t was used, especially in the 13th century, in support of claims of political authority by the papacy.

JSON output

```
{
"item": "Donation of
Constantine",
"alleged_author":
"Constantine the Great",
"alleged_date": "4th
century",
"alleged_location":
"Rome",
"item_type": "decree",
}
```

Listing 6.14: CH Item metadata extraction from source text to structured JSON output

Cognizer Identification

Aim: Identify the subset of Cognizers who make scholarly statements about the given CH items

Input: Cleaned article + item(s) metadata (output of previous step, see Section 6.6.1)

Output: JSON with `is_cognizer` classification, coreferences

This component employs GliNER [225] for NER, targeting people, organizations, groups, and locations. GliNER identifies precise character-level spans compared to LLMs, enabling us to exactly identify and group the paragraphs in which each entity appears. Among the selected paragraphs, we aimed at identifying exclusively the subset of entities who express opinions, omitting those who do not from later steps. We created a prompt which instructs the model to check if the entity extracted from GliNER is expressing an opinion on the CH item(s) extracted in Step 1 (6.6.1). The prompt asks to return a binary classification (is_expressing_opinion: True/False) alongside additional textual mentions and co-references of the given entity. The task relies on a combination of In-Context Learning (ICL) and Chain of Thought (CoT) in a Few-Shot setting (with 3 examples). Listing 6.15 shows a paragraph where **Lorenzo Valla** is mentioned alongside the output.

Input: Source text

Later, the humanist and scholar **Lorenzo Valla** argued in his philological study of the text that the language used in manuscript could not be dated to the 4th century.[21] The language of the text suggests that the manuscript can most likely be dated to the 8th century. **Valla** believed the forgery to be so obvious that he suspected that the Church knew the document to be inauthentic.

JSON output

```
{
"entity": "Lorenzo Valla",
"start": 605,
"end": 618,
"label": "person",
"is_cognizer": true,
"is_subject": true,
"mentions": [
"Lorenzo Valla",
"Valla"
]
}
```

Listing 6.15: Cognizer identification from source text to structured JSON output

Entity Resolution and Linking

Aim: Enrich Cognizers with biographical information; paragraphs grouped by Cognizer

Input: Cognizers and coreferences

Output: JSON with relevant paragraphs grouped by Cognizer and its biographical information

The third component performs coreference resolution and Entity Linking (EL). It clusters entities through identical mentions across paragraphs, then uses the Wikibase API[38] to retrieve candidates for the longest mention of each entity. Each candidate receives a score based on name similarity (Levenshtein distance[39]) between mentions and Wikidata labels/aliases, entity type compatibility, and for people, occupation relevance using Wikidata property `P106`[40] (prioritizing scholarly occupations likely to express opinions in this domain).

Input: Clustered entities

```
{
"primary_mention":
"Lorenzo Valla",
"all_mentions": [
"Lorenzo Valla",
"Valla",
"the humanist"
],
"entity_type": "person",
"paragraphs": [0, 3, 7]
}
```

Output: Wikidata entity

```
{
"wikidata_label": "Lorenzo
Valla",
"wikidata_id": "Q214115",
"occupation": ["humanist",
"philologist"...],
"birth_year": 1407,
"death_year": 1457,
"mentions": ["Lorenzo
Valla", "Valla"]
}
```

Listing 6.16: Entity resolution and linking: from clustered mentions to Wikidata-enriched entities

[38] https://www.mediawiki.org/wiki/Wikibase/API
[39] https://pypi.org/project/python-Levenshtein/
[40] https://www.wikidata.org/wiki/Property:P106

Opinion Extraction and Classification

Aim: Extract the first layer of the Cognizer's opinion

Input: Entity + Wikidata Information (if linked) + paragraphs where entity is mentioned

Output: JSON describing (1) the Cognizer's opinion, (2) their opinion type, and (3) the metadata of the opinion (where, when, provenance)

The fourth component extracts and classifies authenticity opinions based on the identified Cognizers (Section 6.6.1). If the entity has been successfully linked to Wikidata, this information is provided to the model as well.

The extraction process captures the main Core Ontological Pattern of the opinion: the opinion target(s) (which documents or artifacts), opinion types following SEBI classifications (Authentic, Forgery, Formal forgery, Content forgery, Neutral), confidence levels expressed by the Cognizer, temporal contexts (when opinions were expressed), and geographic contexts where relevant. See Listing 6.17 for reference.

Input: Source text

Later, the humanist and scholar **Lorenzo Valla** argued in his philological study of the text that the language used in **the manuscript** could not be dated to the 4th century. The language of the text suggests that the manuscript can most likely be dated to the 8th century. Valla believed the **forgery** to be so obvious that he suspected that the Church knew the document [...]

JSON output: Opinion classification

```
"opinions":
[{
"entity": "Lorenzo Valla",
"subject": "Donation of
Constantine",
"opinion": "Forgery",
"confidence": "High",
"date": "1439-1440"
"location": ""
}]
```

Listing 6.17: JSON output from source text about Cognizer, subject of the opinion, provenance, and assessment

Evidence Mining and Feature Assessment

Aim: Enrich the Cognizer's opinion with supporting evidence

Input: Structured opinions + contextual paragraphs

Output: JSON with supporting evidences and evaluations for each opinion

In the fifth component, the model enriches the basic opinion of the Cognizer with evidences and features being evaluated. Features are organized into three categories following the SEBI ontology: *intrinsic features* (content, language, style, orthography), *extrinsic features* (handwriting, ink, material support, physical characteristics), and *provenance information* (historical context, witness accounts, transmission history). For each feature, the system determines evaluation criteria including consistency (does the feature match the alleged period/author?), presence (is the expected feature present or absent?), completeness (is the feature fully preserved/documented?), reliability (can the feature be trusted as evidence?), and veridicality (does the feature represent authentic information?). Each evaluation receives polarity assignment (positive, negative, neutral evidence) and links to supporting scholarly opinions, creating structured representations of evidence-based reasoning in authenticity assessment. See Listing 6.18 for reference.

Input: Source text

[...] **Reginald Pecocke**, Bishop of Chichester (1450–57), reached a similar conclusion. Among the indications that the Donation is a forgery are its language and the fact that, while certain **imperial-era formulas** are **used in the text**, some of the Latin in the document could not have been written in the 4th century

JSON output: Evidence extraction

```
"evidence_evaluations":
[{
  "evidence":
  "imperial-era formulas",
  "feature": "language",
  "evaluation": "presence",
  "polarity": "positive"
}]
```

Listing 6.18: JSON output of identified evidence with evaluation

Hypothesis Extraction

Aim: Enrich the Cognizer's opinion with hypotheses made on the CH item(s)

Input: Opinions + evidence evaluations + full document context

Output: JSON with hypotheses about document origins, intent, etc.

The final component enriches the output with the Cognizer's hypotheses

on the CH item. The hypotheses can be of four types: *authorship hypotheses* (who actually created items if not alleged authors?), *dating hypotheses* (when were items actually created if not alleged dates?), *location hypotheses* (where were items actually created if not alleged locations?), and *motivation hypotheses* (why were items created or forged?).

The system handles cases where Cognizers accept alleged metadata as authentic as well. For consistency and to avoid negated categories (e.g., "not Constantine"), we include polarity (positive/negative) as a field. See Listing 6.19 for reference.

Input: Source text

Later, the humanist and scholar **Lorenzo Valla** argued in his philological study of the text that the language used in manuscript could not be dated to the 4th century. The language of the text suggests that the manuscript can most likely be dated to the **8th century** [...] Valla further argued that **papal usurpation of temporal power** had corrupted the church, caused the wars of Italy, and reinforced the "overbearing, barbarous, tyrannical priestly domination."

JSON output: Hypothesis extraction

```
"hypotheses":{
"authorship": {
"hypothesis":
"Constantine",
"confidence": "High",
"polarity": "negative"
},
"creation_date": {
"hypothesis": "8th
century",
"confidence": "Medium",
"polarity": "positive"
}...
```

Listing 6.19: Hypothesis extraction from source text to structured alternative theories

6.6.2 Knowledge Graph Generation

The final output is mapped to RDF using the algorithms explained in Sections 6.5.2 and 6.5.3. This subsection showcases the produced Knowledge Graphs in RDF-star format and describes the anatomy of our outputs (specifically, this example was generated by the pipeline using Llama 3.3 70B). Figure 6.20 shows the general structure of a generated KG from the GraphDB interface [143]. Each CH item is represented with both alleged metadata (what the item claims to be) and scholarly assessments, as shown in Listing 6.4. The

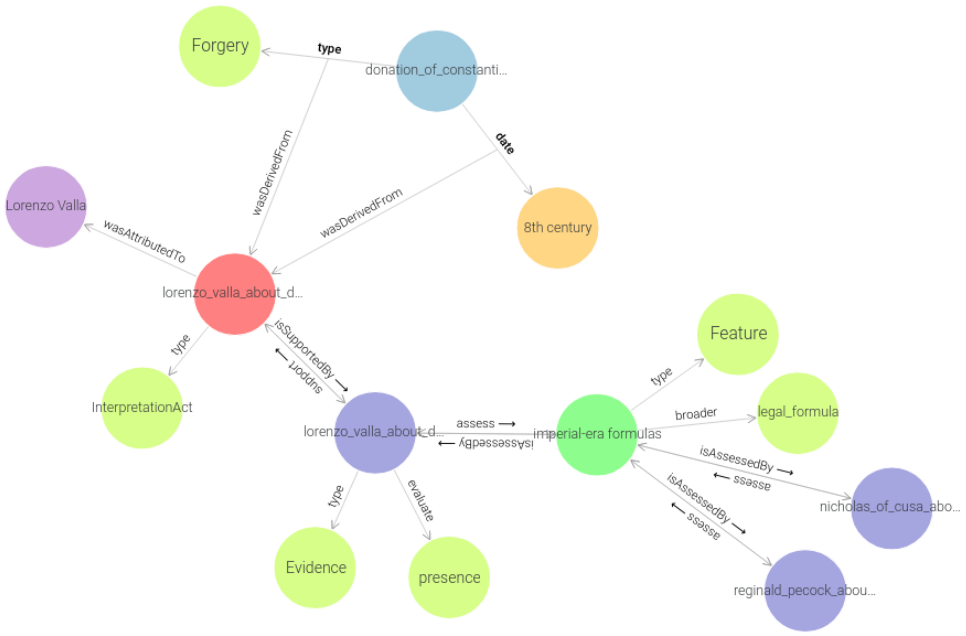


Figure 6.20: Lorenzo Valla's statement about the Donation of Constantine

Donation of Constantine exemplifies this pattern:

Listing 6.4: Document representation with alleged and scholarly metadata

```
# Basic Item information
kb:donation_of_constantine a sebi:Decree ;
    dct:title "Donation of Constantine"@en ;
    dct:coverage kb:rome .

# Item type definition, generated from the text:
sebi:Decree rdfs:subClassOf dcmitype:Text ;
    rdfs:label "decree"@en .

# Alleged metadata as quoted triples (what the item purports to be)
<< kb:donation_of_constantine dct:creator kb:constantine_the_great >>
    prov:wasDerivedFrom kb:donation_of_constantine_self_statement .

```

```
<< kb:donation_of_constantine dct:date kb:constantines_reign_306-337_ad >>
    prov:wasDerivedFrom kb:donation_of_constantine_self_statement .

<< kb:donation_of_constantine dct:coverage kb:rome >>
    prov:wasDerivedFrom kb:donation_of_constantine_self_statement .
```

Listing 6.5 shows Lorenzo Valla's interpretation of the Donation.

Listing 6.5: Lorenzo Valla's interpretation with supporting evidence

```
# Lorenzo Valla as scholarly agent
kb:lorenzo_valla a sebi:Human, dct:Agent ;
    rdfs:label "Lorenzo Valla"@en ;
    owl:sameAs wd:Q214115 ;
    skos:altLabel "Valla"@en ;
    wd:occupation kb:latin_catholic_priest, kb:philologist,
                  kb:philosopher, kb:renaissance_humanist .

# Valla's interpretation act
kb:lorenzo_valla_about_donation_of_constantine a hico:InterpretationAct ;
    sebi:date kb:1439-1440 ;
    prov:wasAttributedTo kb:lorenzo_valla ;
    prov:wasQuotedFrom "donation_of_constantine"^^xsd:anyURI ;
    cito:isSupportedBy kb:lorenzo_valla_about_donation_of_constantine_1 .

# Main authenticity claim
<< kb:donation_of_constantine rdf:type sebi:Forgery >>
    prov:wasDerivedFrom kb:lorenzo_valla_about_donation_of_constantine .

# Alternative dating hypothesis
```

```
<< kb:donation_of_constantine dct:date kb:8th_century >>
    prov:wasDerivedFrom kb:lorenzo_valla_about_donation_of_constantine
     .

# Motivation hypothesis
<< kb:donation_of_constantine sebi:intendedTo kb:political_authority
   >>
    prov:wasDerivedFrom kb:lorenzo_valla_about_donation_of_constantine
     .
```

The supporting evidence for Valla's conclusions is captured through the Evidence graph, shown in Listing 6.6.

Listing 6.6: Lorenzo Valla's philological evidence structure

```
# Evidence node linking feature assessment to interpretation
kb:lorenzo_valla_about_donation_of_constantine_1 a sebi:Evidence ;
    sebi:assess kb:philological_arguments ;
    sebi:evaluate sebi:consistency ;
    sebi:hasEvaluationScore "negative"@en ;
    sebi:support kb:lorenzo_valla_about_donation_of_constantine ;
    ov:confidence 1.0 .

# Feature being assessed
kb:philological_arguments a sebi:Feature ;
    rdfs:label "philological arguments"@en ;
    sebi:isAssessedBy kb:lorenzo_valla_about_donation_of_constantine_1
     ;
    skos:broader kb:language .
```

6.6.3 Evaluation Framework

Our evaluation framework provides a multi-dimensional assessment of the KG generation pipeline, implementing the last task of the ATR4CH methodology as described in Section 5.3.5. We evaluate both the automated extraction

components and the overall discourse representation quality. We integrate human assessment throughout our evaluation pipeline to align KGs, and use F_1 score and G-EVAL metrics. The framework systematically addresses five Evaluation Questions (EQs), aligned with research questions RQ6.2 through RQ6.2 introduced in Section 6.2.

EQ1: CH Item Metadata Extraction Precision: How accurately does the pipeline extract alleged item metadata compared to expert annotations?

Methodology: We formulate this as a multiclass classification task, evaluating the metadata extraction component described in Section 6.6.1 against the Ground Truth (GT). Our classification scheme follows standard evaluation practices:

- **True Positive (TP):** Exact matches between model output and GT
- **False Positive (FP):** Incorrect model predictions
- **True Negative (TN):** Correctly identified absence of metadata when GT is also empty
- **False Negative (FN):** Missing outputs when GT contains valid metadata

To accommodate acceptable semantic variations (e.g., alternative titles, location aliases), we manually review all FP cases to identify outputs that are semantically equivalent to the GT and should be reclassified as TP.

Metrics: We report micro-averaged results for individual metadata categories (Title, Creator, Date, Location) and macro-averaged overall performance using standard precision, recall, and F_1-score calculations.

EQ2: Scholarly Entity Recognition Coverage: How effectively does the entity recognition and opinion frame module identify scholarly agents (Cognizers) present in the source documents?

Methodology: We evaluate the entity extraction component by conducting frequency-based analysis comparing GT entities with model-identified entities as described in Section 6.6.1.

Metrics: We calculate entity-level recall (proportion of GT entities correctly identified) and report the total number of entities detected by the model to assess both coverage and potential over-generation.

EQ3: Evidential Reasoning Extraction Quality: How accurately does the model capture the multi-dimensional evidential reasoning employed by scholars in their interpretations?

Methodology: Given the complex structure of scholarly evidence identified in our ontological framework (Section 6.5.1), where each piece of evidence comprises multiple semantic dimensions (evaluated feature, evaluation perspective, broader feature class, polarity), we implement a custom scoring metric operating on a 4-point scale.

For each evidence prediction, we assign points based on accuracy across these four dimensions, subtracting one point for each incorrectly identified component. This approach accommodates cases where model outputs are semantically similar but not lexically identical to GT annotations.

Example: Consider a scholar arguing that a document is forged due to linguistic anachronisms. If the GT annotation records "lack of regional terms - language - presence - negative" but the model outputs "expected language variety - language - consistency - negative," this represents acceptable semantic alignment despite surface-level differences, warranting partial credit rather than complete penalization.

Score Interpretation:

- **0 points:** Complete extraction failure (equivalent to FN or total FP)
- **1-2 points:** Weak but partially acceptable outputs
- **3-4 points:** Acceptable to strong outputs meeting semantic require-

ments

Scope: Evidence evaluation is restricted to entities successfully matched between model output and GT from EQ2.

EQ4: Hypothesis and Judgment Identification: How accurately does the model extract scholars' interpretative hypotheses and overall authenticity judgments?

Methodology: We apply the same precision, recall, and F_1-score evaluation framework established for EQ1 to assess the hypothesis extraction component described in Section 6.6.1. Model outputs are compared against expert-annotated GT for both specific scholarly hypotheses and overall authenticity determinations.

Scope: Evaluation is limited to the subset of successfully matched entities identified in EQ2 to ensure fair comparison.

EQ5: Overall Discourse Representation Fidelity: Does the complete generated KG provide an adequate representation of the scholarly debate surrounding the CH items' authenticity?

Methodology: To evaluate the representation fidelity, we employ G-EVAL [123]. Since the KGs only represent the opinions inside the text, comparing the source document with a rehydrated version of the KG would heavily bias the evaluation metric. This led us to avoid similarity-based metrics like BLEU, ROUGE, and COMET with the source corpus as used in [70].

We use G-EVAL to evaluate two metrics: *debate correctness* and *debate representativeness*. The first evaluates how well individual scholarly entities and their arguments are represented compared to the GT, penalizing omission of specific entities while rewarding accurate representation of facts, claims, and evidence with proper domain-specific terminology. The second assesses how comprehensively the overall structure and flow of the authenticity debate is captured, including the breadth of scholarly perspectives and their relation-

ships within the discourse narrative.

It should be noted that previous evaluation metrics mostly covered matchable entries between GT and output, whereas G-EVAL evaluates the whole output.

Scope: G-EVAL over rehydrated KGs covers the complete pipeline output against the rehydrated GT.

6.7 Results

This section presents a comprehensive evaluation and preliminary discussion of findings across the five evaluation questions (EQs) outlined in Section 6.6.3. We evaluate Claude Sonnet 3.7, Llama 3.3 70B, and GPT-4o-mini across multiple dimensions of the authenticity debate extraction task (the tables will show only Claude, GPT, Llama for brevity). We begin with simple exploratory SPARQL queries across the 3 KGs and compare the results with the GT, as shown in Listing 6.21.

SPARQL Query for KG Statistics

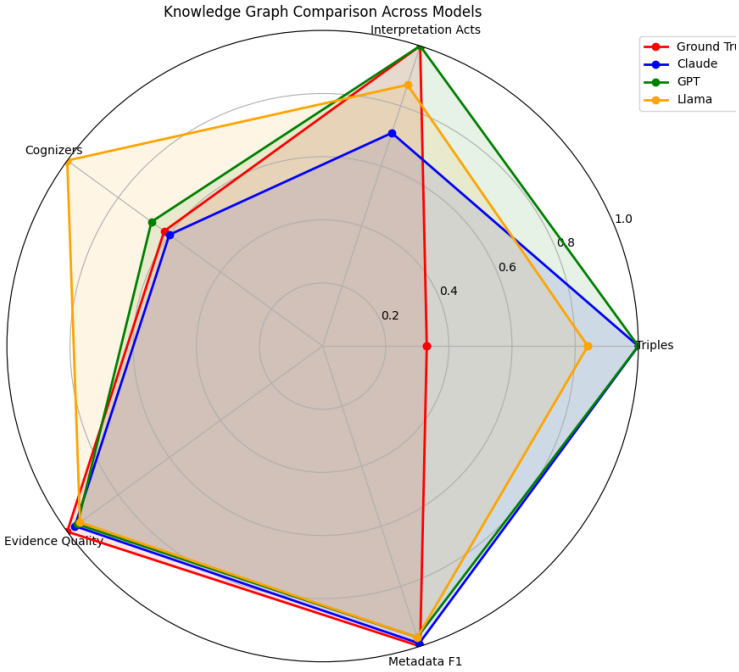
```
SELECT (COUNT(DISTINCT ?entity) AS ?entityCount)
WHERE {
  ?interpretationAct a hico:InterpretationAct .
  ?interpretationAct prov:wasAttributedTo ?entity .
  ?entity a dct:Agent .

  FILTER(!CONTAINS(STR(?interpretationAct),
  "self_statement"))
}
```

Listing 6.21: SPARQL query used to extract entity counts from the KGs for statistical comparison across models

Table 6.3 and Image 6.22 provide an overview of the KGs generated by each model compared to the GT. The models produces more triples than the GT (10,000-12,000 vs. 4,026), primarily because the GT relies heavily on Wikidata entity linking, while the models extract and create explicit triples for information found directly in the text (such as dates, locations, and descriptive

Table 6.3: KE overall metrics

Model	Triples	Interpretation Acts	Cognizers
Ground Truth	4,026	170	164
Claude	10,173	148	103
GPT	12,088	247	201
Llama	10,119	217	172



Listing 6.22: Radar Chart of different KG extractions

metadata). Despite this difference in triple count, the models generate comparable numbers of Interpretation Acts and Cognizers to the GT, suggesting at this stage a similar density of extracted information.

6.7.1 EQ1: CH Item Metadata Extraction Precision

How accurately does the pipeline extract alleged CH Item metadata compared to expert annotations?

As shown in Table 6.4, the performance is high across all models, with F_1-scores ranging from 0.97 to 0.99.

Table 6.4: CH Item Metadata extraction performance across three LLMs

Model	Category	Precision	Recall	F_1-Score
Claude	Titles	1.000	1.000	1.000
	Type	1.000	1.000	1.000
	Creators	0.977	0.977	0.977
	Dates	0.978	1.000	0.989
	Locations	0.978	1.000	0.989
	Overall	**0.987**	**0.995**	**0.991**
GPT	Titles	0.889	1.000	0.941
	Type	0.956	1.000	0.977
	Creators	0.956	1.000	0.977
	Dates	0.911	1.000	0.953
	Locations	1.000	1.000	1.000
	Overall	**0.942**	**1.000**	**0.970**
Llama	Titles	0.933	1.000	0.966
	Type	0.933	1.000	0.966
	Creators	0.933	1.000	0.966
	Dates	0.867	1.000	0.929
	Locations	1.000	1.000	1.000
	Overall	**0.933**	**1.000**	**0.965**

Claude Sonnet 3.7 achieves the highest overall performance with an F_1-score of 0.987. All models show nearly perfect recall, indicating successful extraction of all relevant metadata elements, with precision differences primarily reflecting varying false positive rates. Date extraction shows more variability, with Llama 3.3 achieving the lowest precision (0.867) due to higher FPs rates, as it misclassified the forging date with the alleged dating. For this particular task the challenge was to distinguish between alleged metadata and settled metadata. All models successfully understood the task, showing only small precision drops at varying parameter size.

6.7.2 EQ2: Scholarly Entity Recognition Coverage

How effectively does the entity recognition and opinion frame module identify scholarly agents (Cognizers) present in the source documents? As shown in Table 6.3, the number of Cognizers is relatively similar across models - Table 6.5 shows the number of overlapping entities between

the model's KG and the GT.

Table 6.5: Entity recognition coverage and accuracy

Model	Precision	Recall	F_1	TP	FP	FN
Claude	0.696	0.763	0.728	71	31	22
GPT	**0.718**	**0.912**	**0.803**	145	57	14
Llama	0.626	0.817	0.709	107	64	24

GPT-4o-mini demonstrates superior entity recognition coverage, identifying 77.3% of scholarly agents present in the GT, significantly outperforming Claude (49.5%) and Llama 3.3 (58.8%). It identified the most entities who were expressing opinions. The perfect match rates indicate the proportion of identified entities that exactly match GT annotations. GPT-4o-mini maintains the highest accuracy at 66.0%.

6.7.3 EQ3: Evidential Reasoning Extraction Quality

How accurately does the model capture the multi-dimensional evidential reasoning employed by scholars in their interpretations?

Table 6.6 presents evidence extraction performance using our custom 4-point scoring system that evaluates the accuracy of feature identification, evaluation perspective, feature classification, and polarity assessment.

Table 6.6: Evidence extraction quality and coverage

Model	Mean Score (0-4)	Percentage Score (%)
Claude	**3.87**	**96.8**
GPT-4o-mini	3.84	96.0
Llama 3.3	3.81	95.3

All models demonstrate strong evidence extraction capabilities, with mean accuracies above 0.95%. While GPT-4o-mini achieves the highest precision and recall for entities as shown in 6.5, Claude shows the highest evidence coverage (0.968) in Table 6.6. This pattern highlights that the lower recall in identifying Cognizers by Claude returns in higher precision in downstream tasks.

6.7.4 EQ4: Hypothesis and Judgment Identification

How accurately does the model extract scholars' interpretative hypotheses and overall authenticity judgments?

Table 6.7 presents performance on extracting scholarly hypotheses about items origins and authenticity judgments.

Table 6.7: Hypothesis and judgment extraction performance

Model	Macro F_1	Type F_1	Creator F_1	Date F_1	Location F_1
Claude	0.655	0.652	0.638	**0.791**	**0.923**
GPT	**0.749**	**0.845**	0.484	0.595	0.727
Llama	0.694	0.691	**0.712**	0.762	0.727

GPT-4o-mini achieves the highest overall F_1-score (0.749) for hypothesis extraction, with particularly strong performance in authenticity type classification (0.845). However, the model shows weaker performance in creator hypothesis identification (0.484), suggesting challenges in extracting attribution hypotheses.

Claude demonstrates exceptional performance in geographic hypotheses (0.923 F_1) and temporal hypotheses (0.791 F_1), indicating strength in extracting location and dating alternative theories. Llama 3.3 shows the most balanced performance across hypothesis types, with particularly strong creator hypothesis extraction (0.712 F_1).

The variation across hypothesis types reflects the inherent complexity of scholarly reasoning, with location and date hypotheses generally more explicitly stated than creator attributions or underlying motivations.

6.7.5 EQ5: Overall Discourse Representation Fidelity

Does the complete generated KG provide an adequate representation of the scholarly debate surrounding CH Item authenticity?

The empirical threshold, using the scores produced by G-EVAL on three well-represented articles revised manually (*Posthumous Diary*, *Centiloquium*, *Acámbaro figures*) is set at 0.6-0.7. This result is consistent with other evalua-

tion findings: while the other two models demonstrate higher debate coverage overall, they are penalized for generating more FPs, resulting in lower scores. This evaluation confirms a key pattern in our pipeline - when an entity is correctly identified as a Cognizer, their associated arguments are accurately represented. However, incorrect entity identification leads to error propagation throughout the pipeline, causing the generation of FPs in downstream components. Future iterations of the pipeline should incorporate self-consistency checks at the entity identification stage to reduce error accumulation and improve overall accuracy.

Table 6.8: Per-statement Correctness (G-EVAL scores on 0-1 scale)

Model	Mean	Std Dev	Range
Claude	**0.620**	0.133	0.333 - 0.889
GPT	0.590	0.204	0.222 - 0.889
Llama	0.533	0.153	0.222 - 0.889

Table 6.9: Overall Debate Representativeness (G-EVAL scores on 0-1 scale)

Model	Mean	Std Dev	Range
Claude	**0.607**	0.121	0.333 - 0.889
GPT	0.580	0.199	0.222 - 0.889
Llama	0.523	0.144	0.222 - 0.778

6.8 Discussion and Conclusions

In this section, we discuss overall performance patterns, identified bottlenecks, and possible steps to improve the text-to-KG pipeline while answering our research questions (Section 6.2), followed by our contributions, limitations, and future steps.

6.8.1 Extraction Performance Analysis

To answer RQ1, our evaluation reveals component-specific performance patterns across all tested models. Performance varies significantly across extraction tasks, with all models achieving high scores on metadata extraction (F_1-

scores of 0.965-0.991), moderate performance for entity recognition (F_1: 0.709-0.803), strong evidence extraction capabilities (95.3-96.8% accuracy), and more challenging hypothesis extraction (F_1-scores of 0.655-0.749).

This performance gradient reflects the inherent complexity of different semantic tasks rather than model-specific limitations: extracting alleged metadata proves straightforward across models, while capturing nuanced scholarly hypotheses requires more sophisticated interpretation regardless of architecture. The evidence extraction results demonstrate that contemporary LLMs can effectively capture multi-dimensional evidential reasoning, but they can do so only *when they can identify the Cognizer*—this represents an error propagation problem we identified in the pipeline, as the out-of-GT outputs for evidence extraction are mostly empty or incorrect.

6.8.2 Representation Fidelity and Quality Assessment

To answer RQ2, the generated KGs demonstrate adequate representation of scholarly debate complexity and nuance. While the representation model proves more than adequate as already demonstrated in the BROAST catalogue [147], the *quality* of the automatically generated KGs can still be improved.

G-EVAL scores around 0.6 indicate acceptable discourse representation quality with room for improvement. The successful capture of multi-dimensional evidential reasoning (95.3-96.8% accuracy) shows that LLMs can handle complex semantic relationships, suggesting broader applicability to other humanities domains characterized by multi-perspectival interpretation and evidence-based reasoning. However, the model perspective on specific domain terminology and approaches requires improvement, as the G-EVAL evaluation demonstrates.

6.8.3 Model Comparison and Performance Trade-offs

To answer RQ3, our findings challenge the conventional assumption that larger models always perform better for complex domain tasks. The evaluation reveals distinct performance patterns across models that reflect fundamental

precision-recall trade-offs rather than clear superiority based on parameter count.

Claude 3.7 Sonnet demonstrates lower recall but higher precision, being more conservative in entity classification but achieving greater accuracy in subsequent extraction steps. GPT-4o-mini shows the opposite pattern with higher recall and competitive precision, while Llama 3.3 70B falls between these approaches. Notably, as seen in Table 6.7, GPT-4o-mini performs better since it managed to correctly identify more Cognizers covered in the GT than other models, while having the least parameters of the lot.

The precision-recall trade-off has significant implications for deployment strategies. In production environments where KGs undergo human review and correction, higher recall models may be preferable since updating or deleting erroneous triples is more efficient than creating new KGs from scratch. Conversely, in real-time applications such as RAG systems where extraction occurs without human supervision, higher precision becomes critical to avoid propagating false information.

6.8.4 Methodological Framework Validation

To answer RQ4, our five-step ATR4CH methodology proves effective in coordinating LLM-based extraction with ontological frameworks. Granular evaluation demonstrates that our *divide-and-conquer* methodology enables systematic refinement of individual components while maintaining system coherence. This modular evaluation strategy reveals that different models excel at different subtasks, suggesting potential for hybrid approaches that leverage each model's strengths. The alignment between G-EVAL and other evaluations suggests that self-consistency checks throughout the pipeline (such as prompting models to evaluate their own extraction results) could reduce false positives and false negatives without reducing the necessity of external validation.

6.8.5 Deployment Implications and Cost-Effectiveness

The performance differences between models are relatively modest, while model sizes and costs differ substantially[41]. This suggests that the step-by-step pipeline architecture effectively leverages out the capabilities of different models, making deployment feasible and more cost-effective for CH institutions with varying computational budgets. The competitive performance of different model sizes within sequential pipelines opens two promising research directions. First, fine-tuning approaches could specifically target bottlenecks like Cognizer classification of recognized entities. Second, enhanced pre-processing using specialized tools could filter irrelevant entities before they enter the extraction pipeline. We initially considered frame recognition models for this purpose, but while these models achieve high precision in frame identification, they perform poorly in attribute classification tasks such as identifying Opinion Frame subjects, limiting their utility in entity-oriented pipelines. The methodology's adaptability accommodates diverse institutional landscapes: smaller projects can benefit from intensive human-in-the-loop approaches with API-based models, while larger projects can leverage automated scaling through extensive annotation datasets and local deployment.

6.8.6 Contributions, Limitations and Future Directions

Our primary contributions span three interconnected domains. First, we demonstrated the practical application of the SEBI ontology using RDF-star to represent multi-perspective authenticity claims, enabling structured representation of evidence-based scholarly interpretation while preserving provenance and alternative hypotheses. Second, we introduced a comprehensive five-step methodology for building LLM-centric KE pipelines that addresses the unique challenges of humanities texts through systematic coordination of annotation models, ontological frameworks, and computational tools. The methodology's

[41] As of May 2025, the Claude-3.7-Sonnet API has a cost of $3/million tokens, GPT-4o-mini $0.60/million tokens, and Llama-3.3-70B $0.54/million tokens. The overall cost for 45 articles using the Anthropic API exceeded $20, while for Llama-3.3.-70B and GPT-4o-mini was between $5-10.

technology-agnostic design provides a replicable blueprint adaptable to varying project scales and resource constraints. Third, our technical implementation achieved practical feasibility through a sequential LLM pipeline that successfully captures scholarly reasoning including evidential features, evaluation polarities, and alternative hypotheses. Our approach faces some limitations that can be addressed in future work. The current focus on English Wikipedia sources limits multilingual applicability, particularly important given the *glocal* nature of CH scholarship. Performance on primary scholarly literature remains untested, and two key bottlenecks emerged: Cognizer classification difficulty and dependency on Wikidata linking for optimal performance. Future work will prioritize developing multilingual extraction capabilities, implementing targeted improvements for Cognizer identification through fine-tuning or hybrid approaches, and creating user-friendly tools that enable CH practitioners to customize the extraction process with appropriate human-in-the-loop interfaces. Additionally, working not only with secondary literature but also with primary works from scholars would be a relevant possible contribution. While LLMs show promise for structuring complex scholarly debates, complete automation remains premature, suggesting that balanced human-machine collaboration represents the most viable path forward for knowledge extraction in the Cultural Heritage domain.

Chapter 7

Knowledge Extraction of the *Van den Vos Reynaerde* Authorship Debate

Grimbeert sprac: 'Oem, walschedi?
Of ghi yet wilt, spreect jeghen mi
in Dietsche, dat ict mach verstaen.'[1]

— *Van den vos Reynaerde*, vv. 1457–1459

The methodologies presented in Chapter 4 and Chapter 6 demonstrated the potential of Large Language Models (LLMs) to extract structured Knowledge Graphs (KGs) from concise textual sources. Chapter 6 processed Wikipedia articles summarizing forgery debates, with documents averaging 8,150 characters that already present complex scholarly discourse into accessible encyclopedic summaries. The source texts, while substantive, have a clear advantage compared to primary sources. This chapter addresses a question that emerged from this last implementation: how do LLM-based text-to-

[1]Translation: "Grimbeert said: 'Uncle, do you speak French? If you please, speak Dutch to me, so I may understand."' The joke behind these verses it that, to the common folk in thirteenth-century Flanders, French would have been as incomprehensible as the English expression "double Dutch" ironically suggests today. This passage occurs after Reynaert speaks in pseudo-Latin. Grimbeert's request reflects the class tensions between vernacular Dutch and the languages of power (courtly French and clerical Latin) which served as markers of social hierarchy and instruments for gatekeeping and manipulating knowledge. Text and translation from Besamusca and Bouwman (2009).

KG pipelines perform when confronted with complete scholarly articles rather than summary texts?[2] The transition from encyclopedic summaries to full-length academic publications introduces challenges beyond the mere quantitative scale. Academic articles extend arguments across tens of thousands of words, with interpretations distributed across introduction, literature review, methodology, analysis, and discussion sections. Additionally, there's a different scope. While a summary article usually presents the information assuming the reader does not know the context or the debate in its fullness, a scholarly article is highly contextualized in a scholarship network where highly specialized terminology is used, or references to other articles, theories and interpretations may be more or less implicit. This chapter presents a pipeline employing Retrieval-Augmented Generation (RAG) to extract nanopublications from scholarly articles, following the Digital Hermeneutics model [40]. This proof of concept examines scholarly literature on the Middle Dutch beast epic *Van den vos Reynaerde*, which presents sustained attribution debates about authorship, composition date, and relationship to earlier Reynard cycle traditions. These debates exemplify the interpretative complexity characteristic of literary scholarship, where evidence from linguistic analysis, manuscript tradition, historical context, and intertextual relationships accumulates across argumentative structures that span journal articles, monographs, and edited collections. The chapter contributes to RQ3 (*Can LLM-based extraction systems produce KGs of scholarly interpretations that are sufficiently accurate and complete to answer domain expert competency questions while preserving provenance and epistemic uncertainty?*) by examining an additional aspect of KG extraction about scholarly interpretations. The methodology tests a fundamental hypothesis: targeted content retrieval through RAG can reduce comprehensive scholarly arguments to focused summaries amenable to text-to-KG, without sacrificing the multi-perspectival representation of scholarly

[2]The work presented in this chapter has been submitted to the AIUCD2026 conference as Schimmenti A., van Zundert J., Vitali F., van Erp M. *Scholarly Opinion Mining using LLMs and Knowledge Graphs: the case of the Van den Vos Reynaerde*. It is currently under review.

debate that characterizes interpretation representation like the SEBI ontology. This chapter is organized as follows. Section 7.1 first presents the case study and the context behind the *Van den Vos Reynaerde* authorship debate. Then, it discusses how the Digital Hermeneutics model can be reused. Finally, it details the role of RAG within document-level Knowledge Extraction (KE). Section 7.2 presents the text-to-KG pipeline and how we used Competency Questions to develop the RAG module. Section 7.3 presents the generated nanopublications, the evaluation by the domain expert and listings of the nanopublications. Section 7.4 synthesizes findings and discusses implications for scaling KE to comprehensive scholarly corpora. The full code of the experiment alongside the results are available at the GitHub repository `https://github.com/aschimmenti/digital_hermeneutics_reynaerde.git`.

7.1 Background

This section establishes the theoretical foundations for the RAG-based text-to-KG approach, examining the architectural divergence from the SEBI pipeline and the specific challenges that full-length scholarly articles introduce for ontology-driven knowledge extraction. In addition, it presents the model for knowledge representation adopted when dealing with individual scholarly articles.

7.1.1 Willem die Madocke maecte

Van den vos Reynaerde is a Middle Dutch beast epic dated to the mid-thirteenth century, composed by an author identified through an acrostic signature as *Willem die Madocke maecte*.[3] The work comprises 3,469 verses and constitutes an adaptation of Branch I of the Old French *Roman de Renart*, specifically the section known as *Le Plaid*. The text survives in five manuscript witnesses, with the two complete versions preserved in the Comburg Manuscript (dating between 1380 and 1425) and the Dyck Manuscript[23].

Scholarship on *Van den vos Reynaerde* has generated debates across mul-

[3]Translation: William who made Madocke

tiple dimensions. One of the most debated aspect is the authorship of the poem itself. The debate extends beyond the pseudonymous attribution to Willem, with scholars proposing candidates such as Willem van Boudelo, a Cistercian lay brother, although this hypothesis remains contested within the field [205]. Another point of debate is where did the *Reynaertdichter* (i.e. the author of the *Van den vos Reynaerde*) live or came from - the scholars propose dates around 1250, from the Flemish area. The relationship between the Middle Dutch text and its French source material has prompted extensive analysis regarding the degree of adaptation versus translation, with scholars examining whether the work constitutes a translation, adaptation, or independent reworking of its source [23].

The scholarly corpus addressing *Van den vos Reynaerde* includes monographs, journal articles published in specialized venues such as *Tiecelijn* and *Reinardus: Yearbook of the International Reynard Society*, and critical editions with extensive apparatus. This body of literature demonstrates the sustained scholarly engagement characteristic of contested interpretations in medieval literary studies, where evidence from linguistic analysis, manuscript tradition, historical context, and intertextual relationships accumulates across extended argumentative structures [23].

7.1.2 The Digital Hermeneutics Model

Scholarly knowledge production in the humanities generates claims that require attribution, contextualisation, and provenance tracking at a granular level. Nanopublications [82] package individual assertions as discrete, machine-readable, and citable units of scientific information encoded in RDF. Each nanopublication bundles a core claim with structured metadata documenting its origin and publication context, thereby enabling traceability, attribution, and independent verification at the level of individual statements rather than entire documents (Figure 7.1).

This work adopts the Digital Hermeneutics model developed by Daquino et al., [40], which is based on the nanopublication architecture. The model

```
@prefix swan: <http://swan.mindinformatics.org/ontologies/1.2/pav.owl> .
@prefix cw:   <http://conceptwiki.org/index.php/Concept>.
@prefix swp:  <http://www.w3.org/2004/03/trix/swp-1/>.
@prefix : <http://www.example.org/thisDocument#> .

:G1  = { cw:malaria cw:isTransmittedBy cw:mosquitoes }

:G2  = { :G1 swan:importedBy cw:TextExtractor,
         :G1 swan:createdOn "2009-09-03"^^xsd:date,
         :G1 swan:authoredBy cw:BobSmith }

:G3  = { :G2 ann:assertedBy cw:SomeOrganization }
```

Listing 7.1: Example of a nanopublication about malaria. Source: Groth et al., *The Anatomy of a Nano-publication* [82]

addresses the fundamental challenge in knowledge representation for Cultural Heritage (CH) domains: **scholarly assertions** about artifacts, texts, or historical phenomena **rarely exist as isolated facts**. Rather, such assertions emerge from interpretative processes characterized by uncertain evidence, methodological choices, competing hypotheses, and evolving scholarly consensus. Traditional knowledge representation approaches that model only factual claims often fail to preserve the hermeneutical richness necessary for assessing, comparing, and contextualizing scholarly interpretations.

The Digital Hermeneutics model structures knowledge representation across four distinct layers, each addressing specific aspects of the interpretative process. The Digital Hermeneutics model extends the three layers of the nanopublication by distinguishing between factual background knowledge and interpretative assertions, and by enriching provenance information with hermeneutical context necessary for scholarly assessment.

Layer 0: Factual Data contains background knowledge about artifacts and sources that participants in scholarly discourse assume as established. This layer includes bibliographic metadata describing documents, physical and logical descriptions of cultural objects (manuscript folios, photograph series, architectural elements), and explicit relations between artifacts such as citations or material dependencies. Information in Layer 0 represents what scholars treat as factual within their discourse, even when such "facts" may themselves be contested in other contexts. Layer 0 entities can be organized across multiple

named graphs, enabling modular representation of different artifact types or knowledge domains [40].

Layer 1: Scholarly Assertions captures the content of interpretative claims themselves, corresponding to the assertion graph in the nanopublication structure. This layer is inherently domain-dependent, as the ontological structures necessary to represent art historical attributions differ substantially from those required for philological text criticism or archaeological site interpretation. Layer 1 contains the substantive scholarly claims that form the subjects of hermeneutical analysis. The assertion in Layer 1 is packaged as a named graph that can be referenced and reasoned about as a unit.

Layer 2: Hermeneutical Context provides the evidential and methodological context necessary for assessing scholarly assertions, extending the provenance graph of the nanopublication model with domain-specific hermeneutical information. This layer records the type of interpretative claim (attribution, dating, localization), responsible agents (scholars, institutions), temporal information, source materials consulted, methodological criteria applied, degree of certainty expressed, and relations between competing interpretations. Layer 2 makes explicit the argumentation structure underlying scholarly claims, enabling both human assessment and computational reasoning about interpretation quality. The link between Layer 1 assertions and Layer 0 entities is established through properties such as `hico:isExtractedFrom`, which connects interpretative claims to the sources from which they derive.

Layer 3: Provenance documents the computational processes through which structured knowledge was generated, corresponding to the publication information graph in the nanopublication model. When interpretations are extracted from sources through automatic or semi-automatic methods, Layer 3 records the responsible software agents (`prov:wasAttributedTo`), extraction timestamps (`prov:generatedAtTime`), and procedural details. This layer maintains transparency about the role of computational intermediation in knowledge production, distinguishing between the scholarly provenance cap-

tured in Layer 2 and the technical provenance of the knowledge extraction process itself.

The nanopublication structure provides several advantages for representing scholarly interpretations: it separates concerns across representational layers, enables independent validation of different knowledge types, supports citation of specific scholarly claims through persistent URIs, and facilitates reasoning across heterogeneous sources while preserving provenance. Each complete nanopublication packages a scholarly interpretation (Layer 1) with its hermeneutical context (Layer 2) and technical provenance (Layer 3), while referencing factual entities described in Layer 0.

The application of the Digital Hermeneutics model is exemplified by MIMA (Multi-disciplinary Interpretations model on Manuscript Apparatus),[4] which adapts the base model to represent scholarly analyses of illuminated manuscripts [40]. MIMA addresses a common scenario in manuscript scholarship: multiple scholars from different disciplines (paleography, philology, art history) analyze the same artifact, producing interpretations that may be complementary, independent, or contradictory. In MIMA's implementation, Layer 0 represents manuscripts as `crm:E22_Man-Made_Object` individuals with structural components, Layer 1 captures domain-specific scholarly claims linking physical features to conceptual meanings (such as a paleographer asserting that inscriptional capitals convey monumentality), Layer 2 characterizes each interpretation through `hico:InterpretationAct` individuals specifying analysis type, responsible scholars, methodological criteria, and relations to competing interpretations, and Layer 3 records the extraction and encoding processes.

7.1.3 Beyond Paragraph-level Text-to-KG with RAG

Chapter 6 presented a pipeline to extract scholarly debate and to represent each opinion with the SEBI ontology, which is based as well on the HiCO and PROV ontologies.[5]. As we mentioned above, a limitation of the pipeline

[4] `https://mima-data-model.github.io/mima-documentation/`
[5] `https://valentinapasqual.github.io/sebi/`

is the generalizability to scholarly articles. Retrieval Augmented Generation (RAG) combines retrieval mechanisms with generative language models to ground generated text in retrieved evidence [116]. This architecture addresses a fundamental limitation of purely generative approaches: in e.g. question-answering tasks, LLMs generate text based on learned distributions from training data, potentially producing plausible-sounding but factually incorrect or hallucinated content. RAG mitigates this by explicitly conditioning generation on retrieved passages from a knowledge source, constraining outputs to information grounded in retrievable evidence. The canonical RAG architecture consists of three components: (1) a retriever that identifies relevant passages from a corpus given a query, (2) a reader that generates outputs conditioned on retrieved content, and (3) an embedding model that enables semantic similarity comparison between queries and corpus passages. Implementation typically employs dense vector representations through models such as FAISS [106], which enables efficient nearest-neighbor search over large document collections. Prior work has applied RAG to knowledge extraction tasks, though predominantly in entity-centric rather than ontology-driven contexts. In Chapter 2, we saw how RAG can be applied in tasks beyond question answering, e.g. to improve classification performance in low-resource domains by grounding LLM predictions in retrieved examples [37]. Evaluation of RAG is based on common information retrieval tasks [65], with the most common being ground truth reference, human evaluation with relevance labelling, or LLM-as-a-Judge techniques such as G-Eval [83]. Measures include Precision and Recall at k (P@k, R@k), i.e., precision and recall calculated over the top k, e.g. 10, documents or passages [105].

7.2 Methodology

The methodology presented in this chapter extends the approach developed in Chapter 6, adapting it to the specific requirements of extracting and representing scholarly assertions about the authorship of *Van den vos Reynaerde*. This

work develops both a semantic model for representing authorial arguments and an automated pipeline for extracting such arguments from scholarly discourse.

7.2.1 Corpus Selection and Research Questions

The experiment was conducted under the supervision of a domain expert specializing in medieval Dutch literature and computational literary studies. The expert provided guidance on relevant scholarship and formulated research questions that would capture the epistemological structure of the authorship debate. Five articles (2 in English, 3 in Dutch) were selected. The first three are articles on the authorship debate, representing different perspectives and methodological approaches within the scholarly discourse on *Van den vos Reynaerde* authorship. The last two articles discuss slightly different topics, specifically manuscript tradition and edition history (article 4, [209]) and intertextuality and contemporary context references within Medieval beast poems (article 5, [208]). These two articles will be used as "negative" examples: the pipeline should not be able to generate assertions on the authorship attribution as they discuss different aspects.

1. Bouwman and Besamusca's *Of Reynaert the Fox: Text and facing translation of* Van den vos Reynaerde (2009) [23], which provides a comprehensive edition with critical apparatus addressing authorship alongside textual and translation issues;

2. Peeters' *Historiciteit en chronologie in 'Van den vos Reynaerde'* (1973) [149], which examines historical and chronological evidence to support the identification of Willem van Baudelo as the Reynaertdichter;

3. Van Daele's *De robotfoto van de Reynaertdichter* (2005) [205], which critically evaluates the Willem van Baudelo hypothesis through analysis of Cistercian connections, courtly context, and source material;

4. Wackers' *Wat staat er eigenlijk? Over het editeren van Van den Vos Reynaerde* (2016) [209], which surveys the editorial history and methodological approaches to editing the medieval Dutch Reynaert text;

5. Wackers' *Medieval French and Dutch Renardian Epics: Between Literature and Society* (2000) [208], which examines the relationship between intertextual practices and social contexts across the French and Dutch Reynardian traditions.

These articles were selected to represent the range of interpretative positions, evidentiary strategies, and degrees of epistemic certainty characteristic of the scholarly debate. The domain expert formulated research questions for each article focusing on key aspects of the authorship debate and the interpretative frameworks used to support scholarly claims. The research questions were designed to capture:

1. **Authorship attribution**: Who do the authors identify as the creator of *Van den vos Reynaerde*? What is the relationship between the pseudonymous "Willem" and proposed historical candidates such as Willem van Boudelo?

2. **Temporal localization**: When do the authors situate the creation of the work? What historical period or specific date ranges are proposed?

3. **Spatial localization**: Where do the authors situate the creation of the work? What geographical contexts are invoked?

4. **Thematic and referential content**: What subjects, events, and social contexts do the authors identify as referenced within *Van den vos Reynaerde*? How do interpretations of the work's content (references to feudal systems, legal procedures, courtly life, contemporary political events, Flemish toponyms) inform authorship hypotheses?

5. **Cultural and institutional affiliations**: What social, religious, or institutional contexts do the authors associate with the Reynaertdichter? What role do proposed connections to Cistercian culture, courtly environments, or urban patriciate play in authorship arguments?

6. **Expertise and competencies**: What knowledge domains or linguistic competencies do the authors attribute to the creator? How do claims about expertise in Old French narratives, legal procedures, or specific cultural milieus support authorship hypotheses?

7. **Methodological approaches**: What interpretative criteria and types characterize each scholarly analysis? How do authors combine historical, philological, prosopographical, and sociological methods? What evidentiary strategies (textual analysis, source criticism, contextual analysis, comparative analysis, literary analysis...) are employed?

8. **Epistemic stance**: What degree of certainty do the authors express regarding their claims? How does the scholarly discourse negotiate between hypothesis-based reasoning and literature-based argumentation?

The CQs are be used to evaluate the output of the extraction process as shown in Section 7.3.

7.2.2 Ontological Model for Authorial Attribution

Based on the research questions formulated by the domain expert, CIDOC CRM was selected as the primary ontology for representing scholarly assertions. This choice aligns with the Digital Hermeneutics model's recommendation for Layer 1 representation. The implementation draws primarily on CIDOC CRM[6] for core entity and event modeling, supplemented by HiCO[7] for interpretation representation, PROV-O[8] for provenance, CWRC for certainty,[9] FOAF[10] for agent descriptions, DCTerms,[11] and FABIO[12] for bibliographic metadata. Listing 7.1 presents the complete set of namespace prefixes.

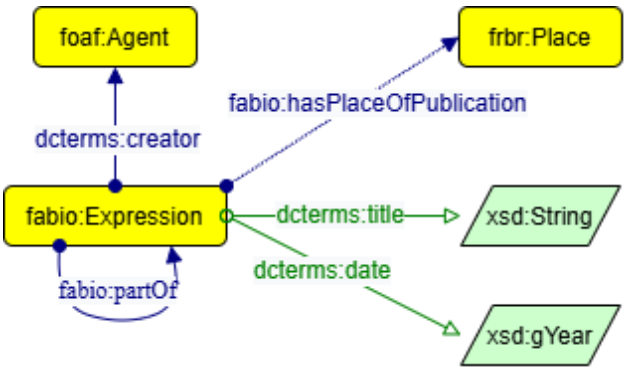
Listing 7.1: Namespace prefixes for the ontological model

[6] https://cidoc-crm.org/
[7] https://marilenadaquino.github.io/hico/
[8] https://www.w3.org/TR/prov-o/
[9] https://sparql.cwrc.ca/ontologies/cwrc-2020-07-14.html
[10] http://xmlns.com/foaf/0.1/
[11] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[12] https://sparontologies.github.io/fabio/current/fabio.html

```
@prefix ex: <http://example.org/> .
@prefix np: <http://www.nanopub.org/nschema#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix fabio: <http://purl.org/spar/fabio/> .
@prefix frbr: <http://purl.org/vocab/frbr/core#> .
@prefix prism: <http://prismstandard.org/namespaces/basic/2.0/> .
@prefix crm: <http://www.CIDOC CRM.org/CIDOC CRM/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix hico: <http://purl.org/emmedi/hico/> .
@prefix cwrc: <http://sparql.cwrc.ca/ontologies/cwrc#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

Layer 0: Factual Representation

Layer 0 represents the factual foundation of the KG, following the Digital Hermeneutics model as illustrated in Figure 7.2. This layer includes document metadata for the scholarly articles and any statements classified as established facts.



Listing 7.2: Factual layer representation (Graffoo diagram)

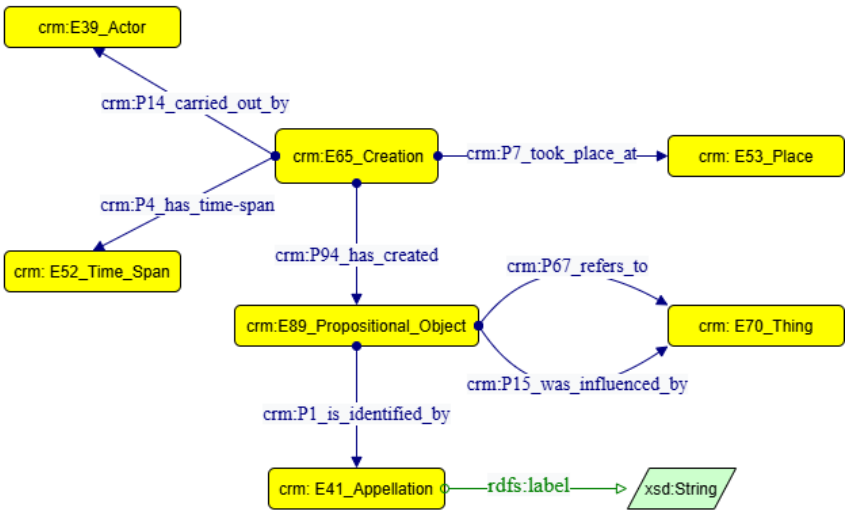
Entity types are mapped to CIDOC CRM classes: `work` entities become instances of `crm:E89_Propositional_Object`, `person` entities become `crm:E21_Person`, `place` entities become `crm:E53_Place`, `organization` entities become `crm:E74_Group`, `date` entities become `crm:E52_Time-Span`, and

`language` entities become `crm:E56_Language`.

Layer 1: Assertion Representation

Layer 1 represents the scholarly assertions about *Van den vos Reynaerde* and its creator. The representation focuses on two primary aspects: the creation event of the work and prosopographical information about the Reynaertdichter.

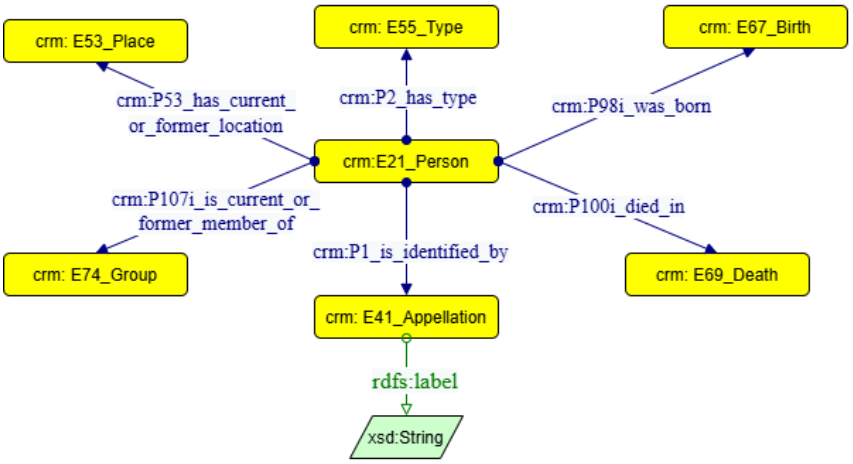
Figure 7.3 illustrates the model for representing the creation event. The central node is an instance of `crm:E65_Creation`, connected to the work (`crm:E89_Propositional_Object`) via `crm:P94_has_created`. The creator is represented as a `crm:E49_Actor`, accommodating both named individuals (such as Willem van Baudelo) and anonymous agents (such as the pseudonymous "Willem"). The event is connected to its creator via `crm:P14_carried_out_by`.



Listing 7.3: Model to represent the *Van den vos Reynaerde* creation event (assertion layer, Graffoo diagram)

Temporal and spatial contexts are attached to the creation event through `crm:P4_has_time-span` and `crm:P7_took_place_at` respectively. Influences on the creation are represented via `crm:P15_was_influenced_by`, which can point to other works, cultural contexts, or individuals. References within the work are captured through `crm:P67_refers_to`, connecting the work to entities mentioned in its discourse.

Figure 7.4 presents the model for representing the Reynaertdichter as a `crm:E21_Person`. This model addresses research questions concerning authorship attribution (CQ 1), cultural and institutional affiliations (CQ 5), and expertise and competencies (CQ 6).



Listing 7.4: Model to represent the Reynaertdichter (assertion layer, Graffoo diagram)

Biographical events are modeled explicitly: birth and death are represented as instances of `crm:E67_Birth` and `crm:E69_Death` respectively, connected to the person via `crm:P98_brought_into_life` and `crm:P100_was_death_of`. Each event can have associated time spans and places.

Several properties require specialized representation strategies. Language competence is modeled by creating a type (subclass of `crm:E55_Type`) for each language, with the person connected via `crm:P2_has_type`. This approach follows recommendations from the CIDOC CRM discussion forum[13]. Expertise domains are similarly modeled by minting types as subclasses of `crm:E55_Type`. Roles and occupations are represented as activities (instances of `crm:E7_Activity`), with the person's participation expressed through `crm:P14_carried_out_by` and the role specified as a type via `crm:P2_has_type`. Group membership is expressed through `crm:P107_has_current_or_`

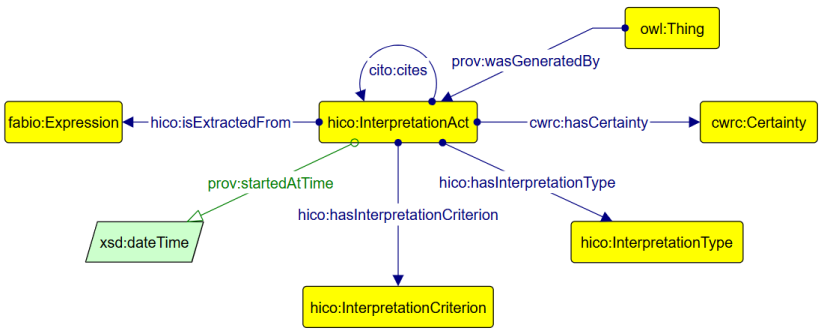
[13] `https://CIDOCCRM.org/Issue/ID-429-p72-has-language`

`former_member`, and residence information uses `crm:P74_has_current_or_former_residence`.

Layer 2: Interpretation Representation

Layer 2 implements the HiCO model for representing the interpretation act itself, as shown in Figure 7.5.[14] This layer addresses research questions concerning methodological approaches (CQ 7) and epistemic stance (CQ 8). Each scholarly article generates an instance of `hico:InterpretationAct` that includes:

- `hico:hasInterpretationType`: The type of interpretation (philological, historical, linguistic, semiotic, paleographic, prosopographic, or sociological interpretation);
- `hico:hasInterpretationCriterion`: The criteria used for the interpretation (diplomatic interpretative transcription, literal transcription, hypothesis-based reasoning, literature-based argumentation, comparative analysis, or authoritative citation);
- `hico:hasCertaintyLevel`: The degree of confidence in the interpretation.



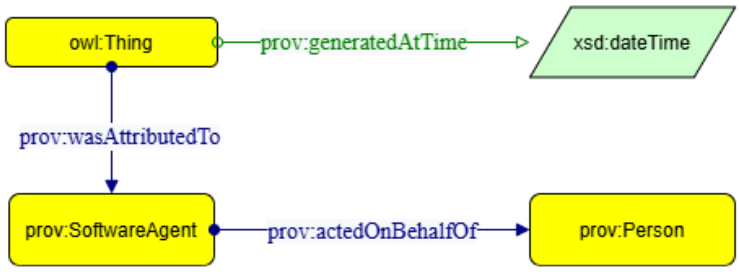
Listing 7.5: Layer 2 - Interpretation layer (Graffoo diagram)

The interpretation act is connected to the graph of assertions through the HiCO model's properties, establishing the epistemological context for the Layer 1 assertions.

[14] The diagram has been adapted from Daquino et al. [40]

Layer 3: Computational Provenance

Layer 3 (Figure 7.6) documents the computational processes through which the structured knowledge was generated. This layer records the software agents responsible for extraction (`prov:wasAttributedTo`), timestamps (`prov:generatedAtTime`), and procedural details of the extraction pipeline. Layer 3 maintains transparency about computational intermediation in knowledge production, distinguishing between the scholarly provenance captured in Layer 2 (the interpretation act by human scholars) and the technical provenance of the knowledge extraction process itself (the automated pipeline).



Listing 7.6: Layer 3 - Provenance layer (Graffoo diagram)

Nanopublication Structure

The complete semantic representation is structured as a nanopublication, following the Nanopub ontology. The Digital Hermeneutics nanopublication consists of four named graphs:

1. **Factual graph**: Contains the Layer 0 statements (entities definitions and factual information);
2. **Assertion graph**: Contains the Layer 1 assertions (authorial arguments);
3. **Provenance graph**: Contains the Layer 2 interpretation metadata and Layer 3 computational provenance;
4. **Publication info graph**: Contains metadata about the nanopublication itself.

The factual layer (Layer 0) is maintained as a separate named graph, allowing facts to be shared between multiple assertions from different scholars. This architecture enables queries that distinguish between consensual facts and contested interpretations.

7.2.3 Pipeline Architecture and Implementation

Having established the ontological model for representing authorial arguments about *Van den vos Reynaerde*, this section describes the automated pipeline that extracts such arguments from scholarly texts and transforms them into the semantic structures defined above. The text-to-KG pipeline consists of four sequential stages: (1) document pre-processing and question answering, (2) entity extraction, (3) relationship extraction, and (4) RDF generation. Each stage transforms the scholarly text progressively from unstructured discourse to structured semantic assertions. The pipeline is a simplified version of the SEBI pipeline presented in Chapter 6 and this test will be used as proof of concept for future implementation steps, following the ATR4CH methodology in its fullness.

Stage 1: Document Preprocessing and Question Answering

The first stage processes scholarly articles to extract relevant information for the predefined research questions. This stage employs a Retrieval-Augmented Generation (RAG) approach to ensure that responses are grounded in the source text.

The document is segmented using a section-aware chunking strategy that preserves structural boundaries and maintains footnote associations. Chunks are created at paragraph boundaries while respecting section divisions, with a target size of 800 tokens and an overlap of 100 tokens between adjacent chunks to maintain contextual continuity.

For the embeddings, the system uses Voyage AI's contextualized embeddings (model: `voyage-context-3`),[15] which preserve document structure by encoding chunks within their section context. The contextualized embedding

[15] `https://docs.voyageai.com/docs/embeddings`

process groups chunks by section and generates embeddings that capture both local and structural information. These embeddings are stored in a FAISS index for semantic retrieval [54].

The question-answering process operates sequentially, processing each research question while maintaining context from previous answers. For each question, the system:

1. Generates a query embedding using the same contextualized embedding model;

2. Retrieves the top-k most relevant chunks from the FAISS index;

3. Re-ranks retrieved chunks using Voyage AI's reranking model (`rerank-2`) to improve precision;

4. Constructs a prompt that includes document metadata, retrieved passages, previously answered questions, and few-shot examples;

5. Generates an answer using a LLM (GPT-4o-mini).[16]

The sequential processing allows subsequent questions to reference information from earlier answers, maintaining coherence across the question set.

Stage 2: Entity Extraction

The entity extraction stage identifies and categorizes entities mentioned in the question-answer pairs using a structured output approach with a predefined JSON schema.

The entity taxonomy encompasses fourteen types aligned with the CIDOC CRM classes defined in the ontological model: `person`, `reference`, `role`, `place`, `work`, `date`, `historical_context`, `organization`, `language`, `methodology`, `theory`, `genre`, `concept`, `event`, `group`, `activity`, `characteristic`, and `material`. The distinction between `person` and `reference` is particularly

[16] `https://platform.openai.com/docs/models/gpt-4o-mini`

significant: entities classified as `person` represent individuals discussed as subjects, while `reference` denotes persons or works cited as supporting evidence by the article's authors.

The extraction process provides each entity with: (1) the entity name as it appears in the text, (2) the entity type from the predefined taxonomy, (3) contextual information indicating where the entity was mentioned, and (4) a confidence score reflecting the certainty of the extraction and classification.

Document metadata (article title, authors, publication date) is explicitly excluded from entity extraction and it is inserted as a fixed input from the previous stage.

Stage 3: Relationship Extraction and Interpretation Metadata

The relationship extraction stage constructs a KG by identifying connections between entities according to the ontological model. This stage operates in three sub-steps: node consolidation, relationship extraction, and interpretation metadata extraction.

First, extracted entities are consolidated into graph nodes, with duplicate entities merged based on name and type similarity. Each node receives a unique identifier and retains the highest confidence score among merged entities.

The relationship extraction step identifies connections between entities using a controlled vocabulary of twenty relation types designed to align with the properties defined in the ontological model. The relation types correspond to the patterns illustrated in Figures 7.3 and 7.4:

- **Creation relations**: `created_by`, `created_during`, `created_in`
- **Spatial and temporal relations**: `located_in_space`, `located_in_time`
- **Influence and reference**: `influenced_by`, `refers_to`
- **Membership and association**: `associated_with`

- **Linguistic properties**: `speaks_language`, `written_in_language`, `has_genre`
- **Biographical properties**: `place_of_birth`, `date_of_birth`, `place_of_death`, `date_of_death`, `lived_in`, `educated_at`
- **Professional and expertise relations**: `has_occupation`

Each relation includes domain and range constraints to ensure type compatibility. A critical aspect of this stage is the assignment of a `claim_type` attribute to each relation, distinguishing between `established_fact` and `authorial_argument`. This distinction enables the separation of Layer 0 (factual layer) from Layer 1 (assertion layer). Facts represent information about which there is scholarly consensus or which can be independently verified, while authorial arguments represent interpretative claims specific to the article being analyzed.

The final sub-step extracts metadata about the interpretation itself, corresponding to Layer 2 of the Digital Hermeneutics model. For each article, the system identifies:

- **Interpretation type**: The analytical approach employed by the scholar (philological, historical, linguistic, semiotic, paleographic, prosopographic, or sociological interpretation);
- **Interpretation criteria**: The evidentiary strategies used (diplomatic interpretative transcription, literal transcription, hypothesis-based reasoning, literature-based argumentation, comparative analysis, or authoritative citation);
- **Certainty level**: The degree of confidence expressed by the scholar in their claims.

This metadata populates the `hico:InterpretationAct` structure illustrated in Figure 7.5.

Stage 4: RDF Generation

The final stage transforms the extracted KG into RDF triples following the ontological model. The implementation generates nanopublications in TriG format, with separate named graphs for facts (Layer 0), assertions (Layer 1), and provenance (Layers 2 and 3).

The RDF generation process adopts an event-centric approach, grouping relations by the events they describe. Relations that contribute to the same event are processed together to mint a single event instance with all relevant properties.

The grouping strategy uses the following event clusters corresponding to the ontological model:

- **Creation cluster**: Groups `created_by`, `created_during`, `created_in` relations to construct a `crm:E65_Creation` event
- **Birth cluster**: Groups `place_of_birth`, `date_of_birth` to construct a `crm:E67_Birth` event
- **Death cluster**: Groups `place_of_death`, `date_of_death` to construct a `crm:E69_Death` event
- **Association cluster**: Processes `associated_with` and `refers_to` relations
- **Influence cluster**: Processes `influenced_by` relations
- **Location clusters**: Groups spatial and temporal location relations

For each cluster, the system mints a unique event URI and attaches all relevant participants, times, and places to this event. This approach ensures that the resulting RDF graph represents a coherent event structure rather than disconnected property assertions.

Table 7.1 summarizes the mapping from schema relations to ontological patterns.

Table 7.1: Mapping from schema relations to CIDOC CRM patterns

Relation	CIDOC CRM Class	Properties
created_by	E65_Creation	P14_carried_out_by
created_during	E65_Creation	P4_has_time-span
created_in	E65_Creation	P7_took_place_at
speaks_language	E55_Type	P2_has_type
has_expertise	E55_Type	P2_has_type
has_role	E7_Activity	P14_carried_out_by, P2_has_type
has_occupation	E7_Activity	P14_carried_out_by, P2_has_type
associated_with	E74_Group or E55_Type	P107_has_current_or_former_member
influenced_by	E65_Creation	P15_was_influenced_by
refers_to	E89_Propositional_Object	P67_refers_to
place_of_birth	E67_Birth	P7_took_place_at, P98_brought_into_life
date_of_birth	E67_Birth	P4_has_time-span, P98_brought_into_life
place_of_death	E69_Death	P7_took_place_at, P100_was_death_of
date_of_death	E69_Death	P4_has_time-span, P100_was_death_of
lived_in	E21_Person	P74_has_current_or_former_residence
educated_at	E7_Activity	P14_carried_out_by, P7_took_place_at

7.3 Results

The text-to-KG pipeline processed five scholarly articles through the RAG-based extraction system, generating nanopublications structured according to the Digital Hermeneutics model. Each nanopublication contains assertion graphs (Layer 1) representing scholarly interpretations about *Van den vos Reynaerde*, hermeneutical context graphs (Layer 2) documenting evidential and methodological information, and factual data graphs (Layer 0) recording background knowledge about persons, places, and textual entities.

The extraction process yielded varying quantities of structured information across the corpus. Table 7.2 presents the triple count for the generated

nanopublications.[17]

Table 7.2: Extraction statistics for nanopublications generated from scholarly articles (temperature 0.3)

Source	Layer 0 Triples	Layer 1 Triples	Layer 2 Triples
Van Daele (2005)	44	17	9
Besamusca and Bouwman (2009)	31	18	17
Peeters (1973)	53	9	6
Wackers (2016)	28	5	7
Wackers et al. (2000)	32	0	8
Total	**188**	**49**	**47**

As for hyperparameters, we started by using a fixed temperature of 0.0 for every single LLM call. The temperature parameter controls the randomness of the model's output: lower values lead to more deterministic and reproducible responses, which is particularly desirable in experimental settings where consistency across runs is required [93]. After some iterations, we tested slightly higher temperature settings, settling empirically at 0.3, exclusively for relationship extraction. This solution proved particularly beneficial, as the number of triples and the variety of statements was positively affected, with almost double the correct statements for two of the three main articles under scrutiny. Each call uses a JSON schema to enforce JSON outputs. The out-of-scope articles correctly produced no assertions, which was surprising as we expected more hallucinated content as a possible drawback. The domain expert evaluated the last of the three iterations with the 0.3 temperature setting. The repository also contains results for the remainder of the iterations.[18]

7.3.1 Authorship Attribution Interpretations

The corpus includes three articles that propose authorship attributions for *Van den vos Reynaerde*. Two articles [205, 149] present the hypothesis that Willem van Boudelo (or Willem van Baudelo) composed the work, while one article

[17] Layer 3 has the same number of triples each time to document the extraction process
[18] `https://github.com/aschimmenti/digital_hermeneutics_reynaerde.git`

[23] identifies the author as Willem without attributing the name to a specific historical individual.

Van Daele (2005) proposes that Willem van Boudelo, identified as a Cistercian *conversus* (lay brother), composed *Van den vos Reynaerde* circa 1250. The assertion graph (see Listing 7.2) contains a `E65_Creation` event with `P14_carried_out_by` linking to Willem van Boudelo, `P4_has_time-span` specifying circa 1250, and `P15_was_influenced_by` documenting Cistercian influence. The graph includes an `E7_Activity` event characterizing Willem as a Cistercian conversi, a `E67_Birth` event situating his birth in the early thirteenth century, and a `E69_Death` event localizing his death to Boudelo. Additionally, the Cistercian order's membership relation to Willem van Boudelo appears through `P107_has_current_or_former_member`. The hermeneutical context documents comparative analysis, hypothesis-based reasoning, and linguistic interpretation as interpretative criteria, with medium certainty.

Listing 7.2: Assertion layer of Van Daele (2005)

```
ex:Cistercian_order crm:P107_has_current_or_former_member
    ex:Willem_van_Boudelo .

ex:Van_den_vos_Reynaerde_Creation a crm:E65_Creation ;
    crm:P14_carried_out_by ex:Willem_van_Boudelo ;
    crm:P15_was_influenced_by ex:Cistercian ;
    crm:P4_has_time-span ex:circa_1250 ;
    crm:P94_has_created ex:Van_den_vos_Reynaerde .

ex:Willem_van_Boudelo_Birth a crm:E67_Birth ;
    crm:P4_has_time-span ex:early_13th_century ;
    crm:P98_brought_into_life ex:Willem_van_Boudelo .

ex:Willem_van_Boudelo_Cistercian_convers_Activity a crm:E7_Activity ;
    crm:P14_carried_out_by ex:Willem_van_Boudelo ;
    crm:P1_is_identified_by
```

```
        ex:Willem_van_Boudelo_Cistercian_convers_Activity_appellation
    .

ex:Willem_van_Boudelo_Death a crm:E69_Death ;
    crm:P100_was_death_of ex:Willem_van_Boudelo ;
    crm:P7_took_place_at ex:Boudelo .

ex:Willem_van_Boudelo_Cistercian_convers_Activity_appellation
    a crm:E41_Appellation ;
    rdfs:label "Cistercian convers" .
```

Besamusca and Bouwman (2009) identify the author as Willem based on the acrostic signature "Willem die Madocke maecte" without attributing this name to a specific historical individual. The assertion graph (see Listing 7.3) documents a `E65_Creation` event with `P14_carried_out_by` linking to a Willem entity, `P15_was_influenced_by` documenting Flemish aristocracy influence, and `P4_has_time-span` situating composition in the thirteenth century. The graph includes referential relations through `P67_refers_to` linking *Van den vos Reynaerde* to the Dampierre family, King Nobel, and Reynaert the Fox. An `E7_Activity` event characterizes Willem's association with the Flemish aristocracy, and a type classification through `P2_has_type` identifies Willem as an Old French speaker. The hermeneutical context documents audience analysis, character analysis, comparative literature, contextualization of feudal society, contextualization of socio-political tensions, geographical assessment, hypothesis-based reasoning, linguistic competence assessment, manuscript evidence, social status assessment, socio-political analysis, and thematic exploration as interpretative criteria, with medium certainty.

Listing 7.3: Assertion layer of Besamusca & Bouwman (2009)

```
ex:Van_den_vos_Reynaerde_Creation a crm:E65_Creation ;
    crm:P14_carried_out_by ex:Willem ;
```

```
    crm:P15_was_influenced_by ex:Flemish_aristocracy ;
    crm:P4_has_time-span ex:thirteenth_century ;
    crm:P94_has_created ex:Van_den_vos_Reynaerde .

ex:Willem_Flemish_aristocracy_Activity a crm:E7_Activity ;
    crm:P14_carried_out_by ex:Willem ;
    crm:P1_is_identified_by
        ex:Willem_Flemish_aristocracy_Activity_appellation .

ex:Old_French_Speaker a crm:E55_Type ;
    crm:P1_is_identified_by ex:Old_French_Speaker_appellation .

ex:Old_French_Speaker_appellation a crm:E41_Appellation ;
    rdfs:label "Old French Speaker" .

ex:Van_den_vos_Reynaerde crm:P67_refers_to ex:Dampierre_family,
        ex:King_Nobel,
        ex:Reynaert_the_Fox .

ex:Willem_Flemish_aristocracy_Activity_appellation a crm:
   E41_Appellation ;
    rdfs:label "Flemish aristocracy" .

ex:Willem crm:P2_has_type ex:Old_French_Speaker .
```

Peeters (1973) proposes that William van Baudelo, identified as a Cistercian cleric from the Land of Waes with a lifespan of 1248-1263, composed *Van den vos Reynaerde* after 1258. The assertion graph (see Listing 7.4) contains a `E65_Creation` event with `P14_carried_out_by` linking to William van Baudelo, `P4_has_time-span` specifying after 1258, and `P15_was_influenced_by` documenting territorial disputes as contextual influence. The graph includes referential relations through `P67_refers_to` linking *Van den vos Rey-*

naerde to Bouchard of Avesnes, the Land of Waes, Margareta of Flanders, and Willem of Dampierre. The hermeneutical context documents hypothesis-based reasoning as the interpretative criterion, with medium certainty.

Listing 7.4: Assertion layer of Peeters (1973)

```
ex:Van_den_vos_Reynaerde_Creation a crm:E65_Creation ;
    crm:P14_carried_out_by ex:William_van_Baudelo ;
    crm:P15_was_influenced_by ex:territorial_disputes ;
    crm:P4_has_time-span ex:after_1258 ;
    crm:P94_has_created ex:Van_den_vos_Reynaerde .

ex:Van_den_vos_Reynaerde crm:P67_refers_to ex:Bouchard_of_Avesnes,
        ex:Land_of_Waes,
        ex:Margareta_of_Flanders,
        ex:Willem_of_Dampierre .
```

Wackers (2016) focuses on analyzing toponymic references and animal allegory patterns within *Van den vos Reynaerde* without proposing authorship attribution. The article examines how place names function within the narrative structure and how animal characters map onto human social hierarchies. The assertion graph documents these analytical observations through referential relations between textual entities and literary devices rather than historical attributions or composition contexts. However, the model produced two false positive references for the work. "Comburg" and "Dyck" are manuscript names referring to two witnesses of *Van den vos Reynaerde*. The RAG system produced the following sentence which is true: "The article mentions entities [...] such as the Comburg and Dyck handwritings, which are referenced as crucial sources in the editing process". However, the final triple in Listing 7.5 is wrong.

Listing 7.5: Assertion layer of Wackers (2016)

```
ex:Van_den_vos_Reynaerde crm:P67_refers_to ex:Comburg, ex:Dyck .
```

The extraction error demonstrates a category confusion between work-level and transmission-level entities. The property `P67_refers_to` indicates

that the content of *Van den vos Reynaerde* references Comburg and Dyck, implying these manuscript designations appear within the narrative itself. The correct ontological structure requires distinguishing between the work as propositional object (`E89_Propositional_Object`) and the manuscripts as physical carriers (`E22_Human-Made_Object`) that transmit the text. The relationship between *Van den vos Reynaerde* and the Comburg and Dyck manuscripts should be modeled through `P128_carries` properties linking the manuscript objects to a textual expression (`E33_Linguistic_Object`). However, these relations were not in scope of the RDF generation step.

Wackers, Block, Dufournet, Goossens, Mann, Pastré, Schouwink, Stephenson, Subrenat, Suomela-Härmä, Daele, and Varty (2000) produced an empty assertion layer. There are triples in the factual layer that refer to the work, Willem and the Flanders without a "created by" statement.

Competency Question Coverage

The evaluation framework assessed whether each KG contained enough information to answer the competency questions defined for the *Van den vos Reynaerde* corpus. Table 7.3 presents the coverage analysis across the five generated nanopublications.

The partial coverage designations for Besamusca and Bouwman reflect structural differences in the KG representation. For authorship attribution, the extraction identifies "Willem" through the acrostic signature without linking to a specific historical individual, contrasting with the definite attributions to Willem van Boudelo in Van Daele and Peeters. The temporal dating receives partial coverage because the thirteenth-century time span provides periodization without the precision of the circa 1250 date proposed by Van Daele or the after 1258 dating specified by Peeters.

For Van Daele, social context emerges through the influence relation (`P15_was_influenced_by`) connecting the Creation event to Cistercian influence, the Cistercian conversi Activity event characterizing Willem's organizational role, and the organizational membership documented through `P107_`

Table 7.3: Competency question coverage across extracted KGs

Competency Question	Van Daele	Besamusca et al.	Peeters	Wackers (2016)	Wackers (2000)
Who is proposed as author?	Yes	Partial	Yes	No	No
When was the work created?	Yes	Partial	Yes	No	No
Where was the work created?	Yes	No	Yes	No	No
What linguistic competencies did the author possess?	No	Yes	No	No	No
What social context is associated with the author?	Yes	Yes	Yes	No	No
What does the text reference?	Yes	Partial	Yes	No	No
What methodologies support the interpretation?	Yes	Yes	Yes	No	No

`has_current_or_former_member` linking the Cistercian order to Willem van Boudelo. Geographic localization is documented through the Death event's location property (`P7_took_place_at`) connecting to Boudelo.

For Besamusca and Bouwman, linguistic competency information appears through the `P2_has_type` relation linking Willem to Old French Speaker (Layer 1). Social context emerges through the influence relation (`P15_was_influenced_by`) connecting the Creation event to Flemish aristocracy and the Flemish aristocracy Activity event characterizing Willem's social association. Historical references appear through the `P67_refers_to` relation linking *Van den vos Reynaerde* to the Dampierre family, King Nobel, and Reynaert the Fox. However, the fact that fictional figures (e.g. King Nobel, Reynaert) are conflated with the Dampierre family, one of the historical inspirations of the poems, make this a partial false positive.

For Peeters, social context is documented through the Cistercian cleric

Activity event (Layer 0) and the territorial disputes influence relation. Historical references appear through the `P67_refers_to` relation linking *Van den vos Reynaerde* to Bouchard of Avesnes, the Land of Waes, Margareta of Flanders, and Willem of Dampierre. Geographic localization appears through the Birth event's location property (`P7_took_place_at`) connecting to the Land of Waes (Layer 0).

The absence of positive responses for Wackers and Wackers, Block, Dufournet, Goossens, Mann, Pastré, Schouwink, Stephenson, Subrenat, Suomela-Härmä, Daele, and Varty reflects articles focused on literary analysis and reception history rather than authorship attribution or composition dating. The extraction methodology correctly identified these topics as outside the competency question scope, resulting in empty assertion graphs (Layer 1 with 0 triples for Wackers, Block, Dufournet, Goossens, Mann, Pastré, Schouwink, Stephenson, Subrenat, Suomela-Härmä, Daele, and Varty) or assertion graphs containing literary analysis patterns rather than authorship claims (Wackers) while maintaining non-empty factual data graphs containing textual and thematic information.

Cross-Article Comparison Capability

The nanopublication framework enables systematic comparison of competing scholarly interpretations. The domain expert evaluated whether structured queries could identify convergence and divergence across articles.

Authorship attribution patterns: Two articles [205, 149] propose Willem van Boudelo (or Willem van Baudelo, representing orthographic variation) as author through `E65_Creation` events with `P14_carried_out_by` linking to Willem van Boudelo or William van Baudelo entities. One article [23] attributes authorship to "Willem" without further identification. The structured representation enables queries distinguishing definite historical attributions from attributions to textual signatures.

Temporal dating granularity: Three temporal patterns emerge across the KGs. Van Daele specifies circa 1250 through a `E52_Time-Span` entity.

Peeters provides after 1258 as temporal specification. Besamusca and Bouwman situates composition in the thirteenth century without decade-level precision. The ontological structure accommodates multiple temporal granularities, enabling comparison of dating precision across interpretations. The temporal difference between the circa 1250 and after 1258 datings represents scholarly disagreement documentable through the structured representation.

Geographic localization strategies: The three authorship-focused articles employ different ontological patterns for geographic localization. Van Daele localizes through the Death event location (Boudelo). Peeters employs `P7_took_place_at` on the Birth event to localize the author's origin to the Land of Waes. Besamusca and Bouwman does not include explicit geographic localization properties in the assertion layer. The framework accommodates both person-centric localization through biographical events and the absence of such localization when articles do not make explicit geographic claims.

Social context modeling: Social context appears through multiple ontological patterns across articles. Van Daele documents social context through the Cistercian influence relation (`P15_was_influenced_by`), activity characterization (`E7_Activity` for Cistercian conversi role), and organizational membership (`P107_has_current_or_former_member`). Besamusca and Bouwman employs an activity event for Flemish aristocracy association and an influence relation documenting Flemish aristocracy impact on creation. Peeters documents territorial disputes as contextual influence through `P15_was_influenced_by`. The framework's accommodation of multiple modeling strategies enables representation of different scholarly approaches to contextualizing authorship.

Error Categories and Limitations

The domain expert identified several categories of extraction challenges:

Temporal expression complexity: Natural language temporal expressions contain nuances that the pipeline did not fully capture. Phrases such as "circa 1250" or "after 1258" require conversion to `E52_Time-Span` entities,

potentially losing uncertainty markers present in source formulations. Future implementations should put more focus on completing this aspect and add additional post-processing steps to ensure correct representation.

Model limitations: The model we presented in Section 7.2.2 is only a preliminary step that still lacks the full range of statements possible within the philological discourse. A full workflow, based on the ATR4CH methodology, will finalize the implementation to also account for additional patterns. One of the model limitations addressed by the authors was the impossibility to distinguish the references - e.g., the fact that the *Van den vos Reynaerde* refers to the Dampierre family is a hermeneutical act and it was correctly placed within the KG, but what exactly within the narrative refers to the Dampierre family was neither represented nor possible to extract from the pipeline as is. This aspect will be one of the focus of further development.

Argumentative structure preservation: Extended scholarly arguments developing across multiple text sections may have their conclusions extracted without complete representation of supporting reasoning. The RAG retrieval process prioritizes relevant passages, which may capture interpretative conclusions while omitting premises or evidential details distributed across non-contiguous text sections. The domain expert noted that this limitation affects the ability to reconstruct complete argumentative chains from KG representations. Further work should address this by extending the steps dedicated to the scholar's reasoning.

Methodological specificity: The hermeneutical context graphs document general analytical categories (comparative analysis, textual analysis, historical analysis, hypothesis-based reasoning) without capturing specific techniques, comparison sets, or procedural details described in source articles. For example, Besamusca and Bouwman employs multiple interpretative criteria including audience analysis, character analysis, and socio-political analysis, but the KG does not specify which textual features were examined or which analytical procedures were applied to these different dimensions. The domain

expert observed that this abstraction level suffices for identifying methodological approaches but limits assessment of analytical rigor.

7.4 Conclusions and Future Work

This chapter examined the challenges of extracting and representing complex scholarly reasoning from long-form philological texts within a text-to-KG framework. The proposed pipeline constitutes an early proof of concept demonstrating that it is possible to convert extensive interpretative discourse into structured, machine-readable nanopublications in the CH domain. Despite its preliminary character, the experiment shows that LLMs can assist in distilling domain-specific scholarship into RDF structures, producing outputs that address most of the defined competency questions. The close collaboration with the domain expert proved extremely valuable to both assess the model strengths and limitations and the pipeline itself within a single workflow.

A central contribution of this work is the integration of a RAG module to improve grounding and explainability. The RAG component proved to be also an important explainability aid, as some errors in the KG could be traced back to the answers and to the following pipeline steps. This facilitated diagnosis and iterative correction within the pipeline development. Additionally, tracing back to the source document is particularly important for philological and historical domains. The RAG module also inserted, through reference to the paragraphs, direct traceability to the original article.

While the model we developed had the main scope of representing the creation of the work and the biography of its creator, it still managed to answer most of the CQs. For instance, the influence relations (`P15_was_influenced_by`) could represent historical influences, other works and entities; the activity graph(s) (`E7_Activity`) could model multiple facets of the creator's life, from its work to its collaborations. This heterogeneity highlights the system's capacity to accommodate diverse scholarly perspectives rather than imposing a single canonical representation of context, while also motivating the develop-

ment of guidelines to preserve ontological consistency.

Several methodological limitations emerged. Temporal expression handling is a primary bottleneck: formulations like "circa 1250" or "after 1258" frequently lose epistemic qualifiers when normalized into `E52_Time-Span` entities. Addressing this will require a temporal parser capable of preserving and representing uncertainty.

Preserving argumentative structure is another unresolved issue; the pipeline often extracts interpretive conclusions without their distributed evidential premises, yielding incomplete reconstructions of reasoning chains. Or, better said, the *chain* of this process is somewhat opaque, as the fact that the creator was e.g. "influenced by the Flemish aristocracy to produce the work" could mean, in different contexts, both the fact that the work was commissioned or that the work's content is inspired by. Future work should include explicit steps to detect argumentative dependencies and to reassemble premises and conclusions that span non-contiguous passages.

The model architecture remains at an initial stage toward a full ATR4CH-based workflow. Expanding its representational scope to capture additional analytical categories and reasoning patterns—such as comparative analysis, audience modeling, and hypothesis testing—will better reflect the diversity of philological methods. Likewise, while the present hermeneutical graphs are useful for identifying methodological tendencies, their abstraction level limits assessment of analytical rigor and the reconstruction of procedural detail.

On the technical side, priority should be given to improving entity extraction (date normalization, FRBR-level differentiation, and entity linking along SEBI-like lines) and to decomposing relationship extraction into smaller, semantically coherent tasks. The effect of decoding temperature also requires systematic evaluation: higher temperatures (for example, 0.3) yielded more contextually nuanced relations at the cost of inter-iteration consistency. Characterizing this trade-off is necessary to balance contextual sensitivity with reproducibility for large-scale use.

An earlier prototype of a RAG-based CIDOC CRM retriever was developed to assist mapping relations to ontology properties. In practice, however, this component proved insufficiently useful in its implemented form and was therefore discarded from the experimental pipeline.[19] Its limitations derived from the retriever's dependence on the raw ontology file without sufficiently contextualized examples or pattern-based guidance. Rather than discard the idea altogether, future work should reconceptualize the mapping stage: integrating a probabilistic top-k ranking of candidate properties, enforcing domain-and-range validation, and enriching the retriever with contextualized usage examples derived from annotated corpora could yield a more reliable, semantically grounded mapping process.

Complementing these technical improvements, a further direction is the development of a human-in-the-loop web application to support supervised resolution across pipeline stages. Such an interface would allow domain experts to inspect retrieved evidence, validate or correct entity and relation candidates, and iteratively refine nanopublications before publication. Embedding expert feedback directly into the workflow would improve data quality, maintain interpretive accountability, and provide an auditable provenance trail linking final KG assertions to the textual sources and supervisory actions that produced them. In practice, a semi-automated annotation and validation environment could accelerate annotation cycles while preserving the disciplinary standards required for humanities scholarship.

In summary, this work demonstrates the feasibility of LLM-assisted pipelines for structured knowledge extraction from long philological texts and points toward a path for maturing the approach. Realizing that path will require focused work on temporal uncertainty representation, argumentative reconstruction, richer mapping strategies, and interfaces that foreground expert supervision. Addressing these challenges will move automated philological KGs closer to representing not only scholarly claims but the evidential and

[19]The beta of the CIDOC-Retriever is available as a GitHub repository at `https://github.com/aschimmenti/cidoc-retriever`

methodological practices by which those claims are constructed.

Chapter 8

Synthetic Training Data Generation for Aspect-Based Sentiment Analysis on Book Reviews

The extraction of scholarly opinions from Cultural Heritage (CH) texts, as demonstrated in Chapter 6 and Chapter 7, requires identifying both the aspects under evaluation and the contextual information around those aspects. This pattern, where specific textual elements are evaluated according to evidential criteria, recurs across diverse CH opinion mining scenarios: art historical assessments of attribution, paleographic analyses of manuscript authenticity, philological evaluations of textual variants, and critical reviews of cultural productions. While our work until now addressed this challenge through ontology-driven extraction pipelines targeting authenticity debates, the fundamental task structure aligns with Aspect-Based Sentiment Analysis (ABSA) frameworks established in computational linguistics [156].

ABSA decomposes opinion extraction into granular subtasks: identifying aspect terms (the specific elements being evaluated), classifying aspect categories (the types of features under consideration), and determining sentiment polarity (the evaluative stance expressed). Translating this framework to CH

contexts introduces domain-specific requirements. First, aspects in literary or cultural reviews often correspond to entities within complex ontological frameworks, as in the case of scholarly opinions. Second, CH opinion mining requires distinguishing explicit aspect mentions from implicit references requiring contextual inference, where long discourse (unlike e.g. reviews) anaphorically refers to previous paragraphs. Third, sentiment in scholarly discourse frequently exhibits greater nuance than the positive/negative/neutral trichotomy employed in product reviews, encompassing mixed evaluations where the same aspect receives both praise and criticism.

The primary obstacle to developing ABSA systems for CH domains remains the scarcity of annotated training data. Existing ABSA datasets concentrate on product reviews [156] and restaurant evaluations, domains where aspect categories (food quality, service, ambiance) differ substantially from those relevant to cultural analysis (narrative structure, historical accuracy, aesthetic merit). Manually annotating thousands of CH opinions according to both aspect categories and ontological entity types demands expertise in both domain knowledge and formal ontology, rendering large-scale annotation economically prohibitive for most research projects.

This chapter addresses the training data bottleneck through LLM-based generation of semi-synthetic ABSA corpora, extending the thesis's investigation into scalable text-to-KG methods for CH contexts. While Chapter 6 demonstrated ontology-driven extraction of scholarly opinions from authenticity assessment debates, the current chapter explores whether LLMs can generate training datasets that support this type of structured opinion extraction. Book reviews serve as a methodologically advantageous testing ground: the task structure mirrors scholarly opinion mining (aspect identification, evaluative stance, evidential features) while benefiting from established ABSA frameworks and computational infrastructure. Success in automatically generating annotated book review datasets would validate a transferable approach for producing training data for more complex CH opinion mining scenarios, where

manual annotation at scale remains infeasible. The approach employs LLMs to generate datasets maintaining linguistic authenticity while ensuring systematic coverage of predefined aspect categories. Crucially, the methodology integrates entity typing based on DOLCE foundational ontology classes during generation, producing annotations that support both traditional ABSA tasks and the ontology-grounded extraction required for CH. The resulting dataset enables fine-tuning smaller language models to perform simultaneous aspect extraction, sentiment classification, and entity typing.

The chapter makes three contributions:

1. A dataset of 10,000 book reviews with annotated aspects, categories and types generated using GPT-4o mini, leveraging data from Wikidata, OpenLibrary, and the INEX Amazon/LibraryThing Book Corpus [111], with types annotated using Text2AMR2FRED (TAF) [71]

2. A comprehensive evaluation of Llama 3.1-Instruct 8B[1] on this dataset, establishing baseline performance metrics for the task

3. A fine-tuned version of Llama 3.1-Instruct 8B that serves as a baseline model for the combined ABSA+ET task, demonstrating the feasibility of this integrated approach

These contributions address an extension of the thesis's third research question: *How can generative AI knowledge extraction methods effectively produce knowledge graphs capable of supporting diverse CH research scenarios?* Specifically, this chapter investigates whether LLM-generated synthetic training data can support the development of smaller, efficient models for ontology-grounded opinion extraction—a crucial capability for scaling the text-to-KG approaches demonstrated in Chapters 6 and 7 across broader CH collections where manual annotation remains prohibitively expensive.

This work represents an initial step toward expanding ABSA application beyond consumer reviews into the more nuanced domain of CH. By integrating

[1] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

ABSA and ET through a single model, we establish a foundation for sophisticated opinion extraction systems capable of processing scholarly discourse on literature, cultural artifacts, and historical contexts.

Section 8.1 discusses related work on ABSA, synthetic dataset generation, entity typing, and LLMs. Section 8.2 presents the data and resources used. Section 8.3 describes the semi-synthetic data generation pipeline and model adaptation. Section 8.4 presents evaluation results, including baseline Llama 3.1 8B Instruct performance and fine-tuned model performance with error analysis. Section 8.5 synthesizes findings and discusses future directions.

8.1 Related Work

The existing literature on ABSA for the CH domain reveals several limitations. Current ABSA annotated datasets for book reviews are notably constrained in size and scope [10], with most containing fewer than 500 annotated samples—insufficient for training robust domain-specific models. While LLMs demonstrate impressive natural language understanding capabilities, there is a scarcity of fine-tuned models specifically adapted for ABSA tasks in specialized domains like literature. Furthermore, the prevailing trend of deploying increasingly larger models (100B+ parameters) raises sustainability concerns and creates accessibility barriers. We aim to understand whether efficiently fine-tuned models (8B parameters) can achieve competitive performance with minimal computational resources—a 4-bit quantized version of our model operates on consumer-grade GPUs with just 4GB RAM, dramatically increasing accessibility for researchers with limited computational resources.

8.1.1 Aspect-Based Sentiment Analysis

Aspect-Based Sentiment Analysis (ABSA), unlike simple sentiment analysis, decomposes opinions into the multiple elements that constitute it [156]:

- **Aspect Terms**: Specific words or phrases that refer to particular features, attributes, or components of the entity being reviewed (e.g., character names like “Leopold Bloom”, stylistic elements like “dense prose”,

or thematic components like "narrative structure")

- **Aspect Categories**: Predefined classes that group aspect terms into coherent semantic categories (e.g., "Leopold Bloom" would belong to the "CHARACTER" category, while "dense prose" might fall under "STYLE")

- **Opinion Expression**: The span containing the words or phrases that convey sentiment or evaluation regarding a specific aspect

- **Sentiment Polarity**: The orientation of the opinion expressed about an aspect, typically classified as positive, negative, or neutral

ABSA can also be adapted to detect the cognizer of the opinion and its targets [227], or to assign sentiment not only to the overall opinion but to the individual aspect [168]. In such cases, the input content would also contain the provenance of the opinion, or it would be a reported, indirect opinion (e.g., "**Valentina** thinks that **Ulysses's prose** is too dense...").

8.1.2 Synthetic Dataset Generation

Data augmentation encompasses a set of techniques used in multiple domains to expand existing datasets for Machine Learning. In Natural Language Processing, techniques such as back translation and synonym replacement have been used to expand parallel corpora [118]. Synthetic Dataset Generation leverages a model, such as an LLM, to train smaller LLMs for specific tasks or under-represented domains and languages [140]. It has also been tested for other under-represented domains and tasks where limitation of annotators, funds, and texts is common, especially in the medical field [34]. Most approaches rely on generating text starting from a single prompt or a few rules [127], but the dataset usually results as unnatural or too homogeneous compared to real data, leading to what has been referred to as model collapse [76].

8.1.3 Entity Typing

In open-world approaches, Entity Typing induction from context is often cast as a Natural Language Inference task [117]. In the Semantic Web realm, a hybrid strategy appears most effective: adopting an open-world (or ultra-fine-grained) approach for identifying types, while employing a closed-world approach for the induction of superclasses, aimed at aligning the extracted vocabulary with existing ontologies. This methodology was central to the 2015 Open Knowledge Extraction (OKE) Challenge [138], and it is also the strategy employed by TAF [71].

As pointed out by Ye et al. [222], the types of entities are already part of NER tools. However, when dealing with specific domains, fine-grained types become crucial especially for ontology or vocabulary alignment [179].

8.1.4 LLMs for Knowledge Extraction

Large Language Models (LLMs) are increasingly recognized as valuable tools for generating Knowledge Graphs (KGs) with an expanding body of research focusing on their application in RDF generative tasks [135], Knowledge Base (KB) enrichment [217], or even writing in RDF syntax [67]. LLMs are considered to perform exceptionally well in SA tasks (especially for binary classification and emotion recognition), even in one-shot or few-shot contexts, but they still struggle, as other architectures like BERT, with ABSA. An additional challenge is evaluating their performance, given that traditional datasets are usually unfit to evaluate a generative approach to the task [226].

8.2 Data and Resources

The dataset used for fine-tuning Llama is available on HuggingFace [176]. The code used to fetch the public data, generate the prompts for the semi-synthetic dataset, annotate the DOLCE types, fine-tune and evaluate the model are available as a GitHub repository.[2]

[2] `https://github.com/aschimmenti/absa_et_book_reviews`

8.2.1 Book Reviews Dataset

As base for the reviews, we used a 10,000 set of reviews from the reviews corpus INEX Amazon/LibraryThing Book Corpus [111].

8.2.2 Structured Data

Wikidata and OpenLibrary were used as source for metadata on the books and for the content of the books themselves. Wikidata was queried using the Wikidata dump.[3] The OpenLibrary is a collaborative digital library project, launched by the Internet Archive. It maintains a comprehensive open database of books, authors, works, and editions, with community-contributed metadata. The OpenLibrary API provides programmatic access to this vast collection, allowing developers to query book information including descriptions, cover images, excerpts, subjects, and bibliographic details. It does contain overlapping information with Wikidata, but also many novel characters, places, themes that are not normally described in Wikidata.[4]

8.2.3 DOLCE

A foundational ontology is a domain-agnostic, upper-layer, formalization of knowledge about fundamental entities, such as *Events*, *Processes*, *Objects*, etc. used to structure in a formal language a certain conceptual view of the world [26]. In our work, the DOLCE foundational ontology [26] provides the conceptual backbone and vocabulary for Entity Typing over KG entities, allowing the development and enhancement of domain-specific ontologies, aligned to its structure. The alignment to DOLCE allows seamless integration of KG model outputs with other ontologies and KGs, adopting the same (or compatible) DOLCE model. TAF integrates DOLCE as a base to perform Entity Typing over unseen classes: this feature is the main inspiration for our approach, starting from the assumption that typing a term with a generic class can be further refined to enrich a LOD vocabulary, or even to match it with an existing one with at least one anchoring point—i.e. the DOLCE class itself.

[3]Download date: 19/02/2025

[4]`https://openlibrary.org/developers/api`

8.2.4 Llama 3.1

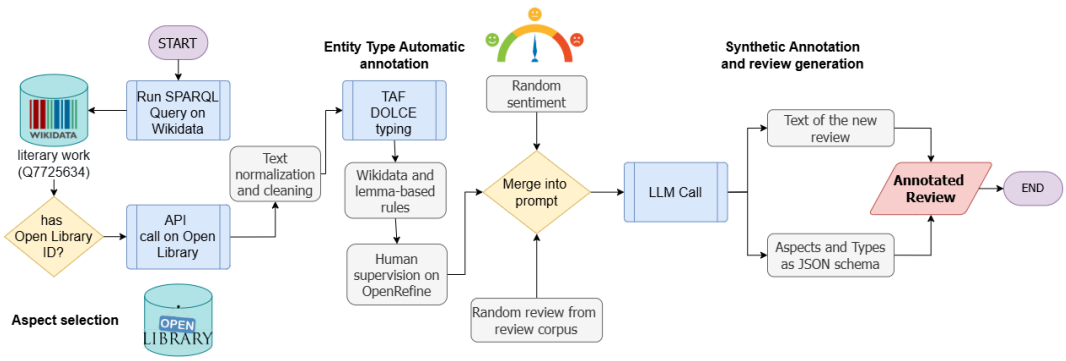
For our baseline implementation, we selected the Llama 3.1 Instruct 8B parameter model based on multiple criteria. Our model selection was guided by three primary considerations: (1) strong performance on the Instruction Following Evaluation (IFEval) benchmark for structured output generation relative to other architectures;[5] (2) relatively low carbon footprint compared to similar models; and (3) seamless integration with contemporary frameworks including Unsloth, Transformers and Ollama. The fine-tuning procedure was implemented using the Unsloth library [87], which provides specialized optimization techniques for LLM adaptation.

8.3 Methodology

In this section, we detail our methodology for the semi-synthetic dataset generation and model fine-tuning.

8.3.1 Semi-synthetic Dataset Generation

We generate our semi-synthetic review dataset in 5 steps as illustrated below. Figure 8.1 shows the flowchart of the semi-synthetic dataset generation process.



Listing 8.1: Synthetic Dataset Generation Pipeline for ABSA-Annotated Book Reviews

[5] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

Data Gathering

The books were sourced from Wikidata[6]: 1,000 instances of literary works were selected. For each, we selected the following properties: P31 (Instance of), P50 (Author), P136 (Genre), P1104 (Pages), P840 (Narrative Location), P674 (Characters), P577 (Publication Date), P1552 (Characteristic), P921 (Main Subject), P180 (Depicts), and P648 (OpenLibrary ID). Thanks to the P31 property, the alignment with DOLCE was immediate through a simple set of rules.

The OpenLibrary API[7] provided additional information such as the description, first sentence, original publication date, subjects, people, locations, time periods, and excerpts. The aspects were unfortunately not as clean (nor already typed) as Wikidata, and had to be extensively cleaned. For this untyped data, we applied TAF.

TAF expects a sentence with at least a verb to perform text-to-graph generation, therefore providing a single word (e.g., "alienation") would not result in a correct output. We therefore elaborate a workaround using the following simple template to return a base classification: ("`<word>` is on the dictionary"). Additional manual cleaning is performed through the tool OpenRefine, with simple multiple macros applied to return the correct types for each term (e.g. the subject key is disambiguated towards genres, people, locations, events etc). Non-English terms were removed.

Aspect Injection

For each book, we randomly selected 1 to 10 aspects following a normal distribution (mean=5, standard deviation=1.5), and to each aspect we assigned a category and a sentiment, distributed randomly as 45% positive, 40% negative and 15% neutral, following the same distribution as the dataset [10]. The aspects were sampled from different categories when available, rather than concentrating on a single aspect type. For each book, 10 reviews were selected

[6] `https://www.wikidata.org/`
[7] `https://openlibrary.org/developers/api`

randomly from a combined pool of Amazon and Goodreads reviews without overlap (i.e., each review was used exactly once as template). This approach maintained linguistic diversity while ensuring consistent sentiment distributions that match real-world book review patterns.

Review Generation

GPT-4o-mini produced the synthetic dataset. The model was instructed to: (1) use the given review as template, (2) to inject the given aspects and sentiments for the new review and (3) to return a JSON with the new review and the annotation.

Aspect Alignments

The selected aspects were aligned with DOLCE using TAF.[8] Given the inconsistency of the tool with single words, we re-aligned the outputs manually with OpenRefine.[9] The final dataset contains 22 types. The high support for InformationEntity is caused by the explicit mention of the book title in the review as an aspect.

Type	Count	Type	Count	Type	Count
InformationEntity	1,830	SocialObject	174	Concept	35
Person	1,174	Collection	146	Activity	34
Event	749	Characteristic	103	Organism	21
Personification	617	Organization	110	Abstract	20
Location	586	PhysicalObject	89	Relation	16
Topic	301	Description	40	System	12
TimeInterval	275			Process	6

Table 8.1: DOLCE type distribution in the dataset

Evaluation

The reviews were evaluated using simple rules, e.g. whether the aspect terms were actually inside the text. GPT-4o-mini was instructed to return both the inserted aspect in the new review and the original input given to add that aspect, to later ensure that the aspect was actually present in the output. To

[8] `https://pypi.org/project/py-amr2fred/`
[9] `https://openrefine.org/`

complement the quantitative evaluation, three evaluators (the author and two co-authors with expertise in computational linguistics and digital humanities) conducted a qualitative assessment of model outputs, examining 100 randomly selected reviews. Each output was manually inspected for aspect identification accuracy, sentiment classification appropriateness, and entity typing correctness. Given the limited sample size, formal inter-rater agreement metrics (e.g., Cohen's kappa) were not computed; however, the evaluators discussed cases of divergent interpretation to reach consensus on the observed patterns. One formatting error was overlooked in 6 out of the 100 reviews, where the aspect term would be returned with the same name as the category (e.g., CONTENT#TOPIC instead of "Civil war") in the annotation (but correct in the review text). A similarity check was used to ensure the original aspect suggested in the prompt was present in the synthetic review. If the review aspect contained the same input, the review was marked as correct.

To illustrate the quality and structure of our generated reviews, we present the following example (for the aspect annotation schema, see Listing 8.2):

> "Reading Ulysses$^{\#\text{TITLE, dul:Inf.Ent.}}$ is like embarking on a labyrinthine journey through Dublin$^{\#\text{LOCATION, dul:Place}}$ with Leopold Bloom$^{\#\text{CHARACTER, dul:Person}}$ as your guide. His character is wonderfully complex, embodying the struggles of everyday life. However, the themes of alienation$^{\#\text{TOPIC, dul:Event}}$ can feel overwhelming, making it hard to connect at times. While it's hailed as high literature, I found the dense prose$^{\#\text{STYLE, dul:Characteristic}}$ a bit off-putting, which might deter casual readers. Despite its accolades, including being listed among the 20th Century's Greatest Hits$^{\#\text{AWARD, dul:SocialObject}}$, I can't help but feel that it sometimes prioritizes style over accessibility. Still, it's a unique experience that challenges conventional storytelling."

In the example, the generated review incorporates various aspects of the book, including character (Leopold Bloom), place (Dublin), themes (alien-

ation), style (dense prose), and award (20th Century's Greatest Hits).

JSON output

```
{
"aspect": "Dublin",
"category":
"CONTENT#SETTING",
"sentiment": "neutral",
"confidence": 0.7,
"mention_type":
"explicit",
"evidence": "labyrinthine
journey through Dublin",
"DOLCEType": "Place"
}
```

Listing 8.2: JSON schema for aspect extraction

8.3.2 Model Adaptation

For the fine-tuning of Llama 3.1-Instruct 8B, we employed the Unsloth library to optimize training efficiency.[10] The training required 1:20:37 hours on an A100 GPU. The model was adapted through Parameter-Efficient Fine-Tuning (PEFT) using LoRA with a rank of 16 and alpha of 16. We trained for a single epoch with a learning rate of `2e-4` using the `AdamW 8-bit` optimizer with weight decay of 0.01 and a linear learning rate scheduler. Training utilized mixed precision (`BF16` where supported) with a per-device batch size of 2 and gradient accumulation steps of 4, effectively creating a batch size of 8 to balance memory constraints with training stability. The training was done on the train split of the dataset (80% train, 20% test).

Each train instance contained system instruction, input and expected output. The system instruction detailed a Chain-of-Thought style description of the task, a detailed description of the JSON schema, and a single example. The training dataset with the full prompt is also available [176].[11] The scripts

[10] `https://unsloth.ai/` Last used: 20 March 2025
[11] `https://huggingface.co/datasets/aschimmenti2/absa_llama"`

to produce the dataset are available on a GitHub repository.[12]

The fine-tuned model is available in three versions through HuggingFace: both a 16-bit and a 4-bit version, as well as only the LoRA adapters [177].

8.4 Evaluation

The evaluation of the fine-tuned model was performed over three iterations and compared with the base Instruction model. The evaluation was performed using the same bit precision (16-bit). Being a generative model, the annotated dataset can only work as Ground Truth (GT). True positives for precision, recall and F_1 score were calculated only on matches between the model's output and the GT. The evaluation was performed three times on the test split of the dataset (2,000 reviews).

Our evaluation reveals that the fine-tuned Llama 3.1 8B model achieves promising performance on the challenging task of literary ABSA with integrated ET. The model demonstrates:

- Strong recall in aspect identification (0.83)
- Competitive overall performance for a relatively small model (7.2 billion parameters)
- High completeness in aspect structure and entity typing (99.39%)
- Particular strengths in identifying character, topic, and author aspects
- Challenges in sentiment classification and implicit aspect recognition

8.4.1 Llama 3.1-8B Instruct Baseline

The base Llama 3.1-8B Instruct model was evaluated on the test dataset to establish a baseline performance. The model demonstrated moderate performance on the ABSA task, with the metrics shown in Table 8.2.

The base model identified a total of 12,653 aspects compared to 6,323 in the ground truth, indicating a tendency toward over-generation (+100.11%

[12] `https://github.com/aschimmenti/absa_et_book_reviews`

Overall Statistics			Performance Metrics			
Metric	**Value**	**%**	**Eval. Type**	**Prec.**	**Rec.**	**F_1**
GT Aspects	6,323	100.00	Aspect	0.3378	0.6759	0.4505
Predicted Aspects	12,653	200.11	Aspect+Sent.	0.2690	0.5384	0.3588
Aspect Matches	4,274	67.59				
Full Matches	3,404	53.84				

Table 8.2: Llama 3.1-8B Instruct Performance Metrics. The Predicted Aspects percentage (200.11%) indicates that the model generated approximately twice as many aspects as exist in the ground truth

more aspects). Despite this, it achieved a recall of 0.67 for aspect identification, meaning it successfully captured approximately two-thirds of the ground truth aspects. However, the precision was notably lower at 0.33, reflecting that many generated aspects did not match the ground truth.

When considering both aspect identification and sentiment classification together, performance decreased significantly, with the F_1 score dropping from 0.45 to 0.36. This suggests that even when the model correctly identified an aspect, it often assigned incorrect sentiment, highlighting sentiment classification as a particular challenge for the base Instruct model.

8.4.2 Llama 3.1-8B ABSA+ET Fine-Tuned Model

Table 8.3 shows a comparable set of metrics to the baseline. Immediately clear is that precision, recall and F_1 score are higher, alongside a higher number of matches, while also having a lower number of Predicted Aspects (from 200.11% to 131%). Table 8.4 shows the distribution between the Fine-Tuned Model and the test dataset. Table 8.5 shows the top distributions of the aspects.

The model demonstrates high recall in aspect identification (0.8342), indicating effective coverage of relevant aspects in the text. The precision of 0.6351 reflects that approximately 36.49% of the model's predicted aspects were not directly aligned with the GT. Considering both entity identification and sentiment classification (full matching), performance increases to an F_1 score of 0.5686.

Overall Statistics			Performance Metrics			
Metric	**Value**	**%**	**Eval. Type**	**Prec.**	**Rec.**	**F_1**
GT Aspects	6,323	100.00	Aspect	0.6351	0.8342	0.7211
Predicted Aspects	8,305	131.30	Aspect+Sentiment	0.5007	0.6577	0.5686
Aspect Matches	5,274	83.42				
Full Matches	4,158	65.77				

Table 8.3: Llama 3.1-8B ABSA+ET Performance Metrics. The Predicted Aspects percentage (131.30%) indicates that the fine-tuned model generated about 31% more aspects than in the GT, showing improved precision compared to the baseline Instruct model

Overall Statistics			Key Differences		
Metric	**Model**	**GT**	**Category**	**Model**	**GT**
Total #aspects	8,305	6,323	BOOK#TITLE	13.31	17.29
Avg. per response	4.16	3.17	CONTENT#CHARACTER	14.05	11.40
Complete aspects	99.39	100.00	BOOK#AUTHOR	5.51	2.12
Sentiment Distribution			**Mention Type**		
Positive	45.55	44.47	Explicit	90.88	83.82
Negative	36.14	40.04	Implicit	9.12	16.18
Neutral	18.31	15.48			

Table 8.4: Fine-Tuned Model Performance Summary

Error Analysis

The errors of the model are the following:

- **Missed Aspects**: 1,048 ground truth aspects (16.58%) went unidentified by the model
- **Incorrect Aspects**: 3,030 predicted aspects (36.49%) did not match ground truth annotations
- **Sentiment Errors**: 1,115 instances (21.15% of matched aspects) where the aspect was correctly identified but assigned an incorrect sentiment

As shown in Tables 8.4 and 8.5, the model's distributional predictions closely mirror ground truth in several categories while showing notable divergences in others. The model identifies 31.3% more aspects overall (8,305 vs. 6,323), suggesting a slightly more fine-grained aspect identification, but not as much prone to over generation as the baseline (12,653).

Top Categories (%)			Top Aspect Types (%)		
Category	**Model**	**GT**	**Type**	**Model**	**GT**
CONTENT#TOPIC	29.87	28.28	InformationEntity	21.61	29.75
CONTENT#SETTING	15.04	15.78	Person	19.33	13.49
CONTENT#CHARACTER	14.05	11.40	Location	13.94	12.98
BOOK#TITLE	13.31	17.29	Topic	11.77	4.52
CONTENT#GENRE	7.23	7.81	Event	11.32	7.37
CONTENT#PERIOD	6.33	9.22	TimeInterval	8.08	7.04
BOOK#AUTHOR	5.51	2.12	SocialObject	4.00	3.37
CONTENT#EVENT	3.57	4.48	Personification	3.54	4.81

Table 8.5: Distribution Comparison Between Model and Ground Truth

Category and Type Performance

For category detection, the model shows particular strength in identifying Characters (+2.65%), Topics (+1.59%), and comments on Authors (+3.39%), while demonstrating comparative weakness in detecting Titles (-3.98%) and Time periods (-2.89%). This pattern suggests that the model has developed stronger sensitivity to discernible narrative elements centered around agents (characters, authors) and thematic content than to structural or temporal elements. The distribution is reflected on the training data, where these aspects were generally less.

In aspect type detection, the model shows notable divergence from ground truth in several DOLCE classes. The model identifies fewer InformationEntity instances (-8.14%) while detecting more Person (+5.84%) and Topic (+7.25%) classifications. This skew toward agentive and thematic elements aligns with the previously observed category detection patterns.

Sentiment and Mention Type Analysis

The sentiment distribution reveals a tendency toward more positive (+1.08%) and neutral (+2.83%) classifications with correspondingly fewer negative assessments (-3.90%).

The most significant distributional difference appears in mention type recognition, where the model heavily favors explicit mentions (+7.06%) while struggling with implicit references (-7.06%). This suggests limitations in the

model's ability to recognize aspects that require deeper contextual inference or domain knowledge.

While the raw metrics might initially appear modest, particularly for full matching (F_1=0.5686), several factors warrant consideration when interpreting these results:

Benchmark Context

- SemEval ABSA challenges for restaurants and laptops typically report F_1 scores between 0.65-0.75 for aspect identification and 0.55-0.65 for aspect+sentiment classification among top-performing systems

- Given the higher complexity of literary reviews and the use of a relatively small model (Llama 3.1 8B), our performance (0.72 for aspect identification) is competitive relative to domain difficulty

Model Behavior Analysis The error analysis reveals important patterns in model behavior:

- **High Recall**: The model's stronger recall (0.83) relative to precision (0.66) indicates a bias toward comprehensiveness over selectivity in aspect identification

- **Sentiment Challenge**: The substantial drop in performance when adding sentiment classification (F_1 from 0.72 to 0.57) highlights sentiment assignment as a primary challenge. Additional analysis of the synthetic dataset and evaluation on other dataset are needed to contextualize this score

- **Entity Focus**: The model's stronger performance on character/person and topic aspects suggests particular sensitivity to these literary elements, which are more discernible than aspects such as Topics, Characteristics and other DOLCE-relevant entities

Qualitative Analysis To complement the quantitative evaluation, we conducted a qualitative assessment of model outputs, examining 50 randomly selected reviews. Several patterns emerged:

- The model excels at identifying explicitly mentioned book elements, particularly characters and narrative settings
- Sentiment classification errors often occur with mixed or nuanced expressions, where positive and negative elements are combined
- The model occasionally replaces the aspect term with the category class if the aspect is implicit, suggesting some challenges with NLU

Entity Typing Performance The integration of DOLCE ontology-based entity typing represents a novel contribution of our approach. The model achieves 99.39% completeness in aspect structure, with only 51 instances missing aspect_type/DOLCEType assignments. This high completeness demonstrates the effectiveness of our approach in simultaneously performing ABSA and ET.

While the distribution of predicted entity types differs from GT in several categories, the model successfully captures the fundamental ontological distinctions in the majority of cases. The confusion between closely related types (e.g., between InformationEntity and Topic) reflects genuine ontological ambiguity in the literary domain.

8.5 Conclusions and Future Work

In this chapter, we presented three main contributions using book reviews as a test case for developing opinion mining infrastructure applicable to CH scholarly discourse: (1) a semi-synthetic dataset of 10,000 book reviews with aspects typed according to DOLCE ontology classes, (2) a comprehensive evaluation of Llama 3.1-Instruct 8B on this dataset, and (3) a fine-tuned model that simultaneously performs ABSA and Entity Typing.

This work validates the feasibility of generating synthetic training data for opinion mining tasks in domains where annotated corpora are scarce. By in-

troducing a semantically rich pipeline to generate synthetic reviews integrating Entity Typing with ABSA, we demonstrated that LLM-based data generation can produce datasets suitable for fine-tuning smaller models while maintaining ontological alignment necessary for KG integration. Book reviews served as an accessible domain for testing this methodology, sharing structural similarities with scholarly opinions –evaluative discourse targeting specific aspects of cultural objects – while avoiding the complexity of acquiring and annotating authentic scholarly texts.

The performance of our fine-tuned model (F_1=0.72 for aspect identification, F_1=0.57 for full matching) demonstrates the viability of this approach, especially considering the relatively small model size (8B parameters). The model's strong recall (0.83) indicates effective coverage of relevant aspects, while its precision (0.64) reflects the challenges of defining exact aspect boundaries in nuanced contexts. Error analysis revealed systematic patterns: the model shows particular strength with explicit mentions of agentive elements (characters, authors) while struggling with implicit references and temporal aspects. Sentiment classification remains a significant challenge for smaller LLMs, especially for aspects with mixed or nuanced sentiment expressions.

The primary contribution of this work extends beyond the book review domain. Having established that semi-synthetic data generation with ontology-aligned entity typing produces functional training datasets, the methodology can now be adapted to generate datasets for scholarly opinion mining in CH contexts. The successful integration of DOLCE typing within ABSA annotations provides the foundation for developing training corpora targeting the opinion mining pipelines presented in Chapters 6 and 7.

Building on these findings, future work should prioritize:

- **CH Scholarly Opinion Datasets**: Applying the semi-synthetic generation methodology to create annotated datasets for authenticity assessment debates, attribution disputes, and other scholarly discourse domains examined in this thesis. This requires adapting the genera-

tion pipeline to handle third-person opinion reporting ("Scholar X argues that..."), evidential reasoning structures, and domain-specific aspect categories derived from CH ontologies

- **Domain Transfer Validation**: Evaluating whether models fine-tuned on book reviews transfer effectively to scholarly texts, or whether domain-specific training data proves necessary. This would establish whether book reviews serve as effective proxy training data or if investment in CH-specific annotation is required

- **Integration with Opinion Mining Pipelines**: Developing methods to incorporate ABSA+ET models as components within the opinion mining architectures developed in previous chapters, particularly for aspect identification and cognizer detection tasks that currently rely on few-shot prompting

- **Model Scaling and Architecture Exploration**: Evaluating larger models in the Llama family (70B+) and alternative architectures (Gemma, Deepseek, Mistral) to determine whether increased parameter count addresses the precision and sentiment classification challenges identified, particularly for implicit aspect recognition and mixed sentiment expressions

- **Knowledge Graph Integration**: Developing methods to automatically integrate ABSA+ET outputs with existing KGs, leveraging the DOLCE ontology alignment for seamless knowledge fusion. This includes not only using DOLCE classes as types but also generating subclasses automatically, following the OKE approach [138]

The demonstrated feasibility of generating ontology-aligned training data for opinion-centric tasks provides a foundation for scaling opinion mining infrastructure across CH domains. By integrating aspect-level sentiment analysis with entity typing grounded in foundational ontologies, this approach supports

the construction of semantically rich KG representing scholarly interpretation, enabling applications ranging from Digital Humanities research to automated extraction of evidential reasoning from cultural discourse.

Conclusions

This dissertation examined the ontology-driven **text-to-KG** task in the Cultural Heritage (CH) domain, addressing a structural tension where traditional supervised learning approaches proved impractical due to limited annotated data. The research responded to a fundamental challenge identified in the Introduction: while digitization efforts have successfully published CH metadata through Semantic Web (SW) technologies, the documents' content remains largely inaccessible through semantic querying [78, 38]. Manual curation scales poorly to the volume of unstructured text in CH collections, yet automatic approaches are hindered by data scarcity when attempting to train supervised models, and by domain-specific challenges that complicate direct application of general-purpose Natural Language processing (NLP) tools. The text-to-KG task itself is not novel; multiple implementations exist in general and specialized domains, as documented in Chapter 1 and Chapter 2. However, its practical deployment encounters implementation barriers that prevent widespread adoption in CH contexts. The sequential subtasks within traditional pipeline architectures—Named Entity Recognition (NER), Entity Linking, relationship extraction, and graph resolution—require component-level accuracy to maintain overall system performance. This dependency chain creates brittleness: errors propagate through the pipeline, and individual component failures can compromise extraction completeness. Furthermore, each subtask typically requires substantial training data for supervised approaches, but CH domains characteristically lack large-scale annotated corpora. This dissertation investigated the integration of LLMs within traditional text-to-

KG pipeline architectures based on two factors: first, LLMs enable few-shot, transfer learning, and fine-tuning approaches that bypass the requirement for extensive annotated training data; second, LLMs demonstrate state-of-the-art language understanding capabilities, potentially enabling structured extraction from complex scholarly discourse that embeds factual claims within interpretative commentary, employs domain-specific terminology, and represents epistemic uncertainty. This investigation does not claim to resolve the fundamental challenges of text-to-KG extraction: domain adaptation, entity disambiguation, and relation extraction remain open research problems. Rather, it demonstrates that LLMs can address specific bottlenecks in CH implementations where data scarcity and linguistic complexity previously prevented the deployment of traditional methods. The investigation of LLM-based extraction revealed two methodological gaps in the existing literature. As discussed in Chapter 5, there was no systematic framework capable of coordinating annotation schema development, ontology alignment, and extraction pipeline design within Cultural Heritage contexts. Furthermore, the evaluation of generative extraction systems proved inadequate when relying exclusively on standard NLP metrics designed for discriminative models. Generative pipelines often produced outputs that diverged from surface-level pattern matching with the ground truth while remaining semantically valid, thus requiring evaluation frameworks that measured functional adequacy for scholarly queries rather than exact lexical correspondence. The chapters of this dissertation collectively addressed these gaps through a progressive methodological and empirical investigation. Chapter 1 defined the text-to-KG task and established its theoretical foundations, distinguishing between Open and Closed Knowledge Extraction paradigms and outlining the architecture of traditional pipelines. Chapter 2 analyzed eleven Cultural Heritage projects that implemented text-to-KG methodologies between 2015 and 2025, revealing that most employed variants of the traditional pipeline. The accompanying literature review of 227 articles published between 2020 and 2025 identified persistent bottlenecks

in Named Entity Recognition, Relationship Extraction, and Entity Linking, while showing that task-specific models generally outperformed LLMs when sufficient training data were available, whereas in data-scarce domains LLMs served as effective facilitators enabling text-to-KG processing. Chapter 5 responded to these findings by introducing *Adaptive Text-to-KG for Cultural Heritage* (ATR4CH), a systematic five-step methodology that integrated LLMs into traditional pipelines through the coordinated use of ontology analysis, Competency Questions, ground-truth annotation, text-to-KG transformation, and multi-layer evaluation. Subsequent chapters demonstrated the implementation of ATR4CH in diverse case studies: Chapter 6 applied the methodology to debates on the authenticity assessment of Cultural Heritage items using the SEBI ontology; Chapter 4 explored its application to archival finding aids through the RiC ontology; and Chapter 7 extended the approach to full scholarly articles, providing a complementary proof of concept. Finally, Chapter 8 presented a further validation through the generation of synthetic training data for domain-specific Aspect-Based Sentiment Analysis (ABSA) in book reviews, demonstrating that LLMs could produce large-scale annotation for domains lacking training resources. Together, these chapters provided a coherent methodological framework and a set of empirical validations that addressed both the absence of integrated approaches to ontology-driven extraction and the need for evaluation models suited to generative systems in Cultural Heritage research. The dissertation addressed three main RQs established in the Introduction, each examined through multiple chapters combining literature analysis, methodological development and empirical validation.

> **RQ1: State of Text-to-KG in Cultural Heritage**
>
> What methodologies have Cultural Heritage projects employed to generate Knowledge Graphs from text, and what domain-specific challenges have these representative projects faced?

Chapter 2 addressed this question through a systematic analysis of eleven

representative projects implementing text-to-KG methodologies between 2015–2025, complemented by literature analysis of 227 papers published from 2020 to 2025. The survey identified projects spanning diverse CH subdomains: biographical heritage (BiographySampo, InTaVia), parliamentary records (ParliamentSampo), military history (WarSampo), olfactory heritage (Odeuropa), musical encounters (MMKG, MusicBO), art history (Viewsari), scientific correspondence (Bernoulli-Euler Digital), and legal documentation (Notarypedia, HyperReal). Seven of the eleven projects employ variations of the traditional pipeline architecture described in Section 2.6.1, decomposing extraction into sequential subtasks. This finding informed subsequent methodological development: rather than proposing novel architectures, effective CH text-to-KG solutions require adapting and orchestrating existing pipeline components to address domain-specific challenges. These challenges where categorized as:

- **Document-level challenges:**
  - **OCR quality degradation**, affecting projects such as WarSampo, MusicBO, and Bernoulli-Euler;
  - **Historical language variation**, complicating processing in Notarypedia and Bernoulli-Euler;
  - **Semantic change across temporal contexts**, impacting the extraction of olfactory terminology, such as in Notarypedia and in Odeuropa.
  - **Coreference resolution** is often addressed as a challenge but, depending on the document's structure, can be circumvented. Paragraph level KE (such as in Odeuropa) does not require explicit resolution; in other cases, extracted data can be "harmonized" through chains as in the MEETUPS case.
- **Component-level challenges:** Concentrated in **Named Entity Recognition (NER)**, **Relationship Extraction**, and **Entity Linking**, with the latter emerging as the most recurrent bottleneck. Entity

Linking difficulties primarily concern **long-tail entities** absent from Wikipedia-derived knowledge bases such as DBpedia and Wikidata.

Data alignment issues manifest across three dimensions identified by [95]:

- **T-Box alignment** requiring schema extension for novel entity types;
- **A-Box alignment** demanding deduplication and consistency checking;
- **URI alignment** involving entity resolution across sources.

The literature analysis revealed concentration in three technological families addressing these challenges: **Large Language Models (44 of 227 papers)**, **Knowledge Graph embeddings (11 papers)**, and **Graph Neural Networks (9 papers)**, with publication activity accelerating from 2023 to 2025 in parallel with surveyed project timelines. These trends address specific pipeline bottlenecks: generalist NER tools such as **GLiNER** [225] enable domain adaptation without extensive retraining; automatic dataset generation supports specialized entity types; LLM-based unsupervised relation extraction leverages extended context windows; and unseen-entity linking approaches such as **KG-ZESHEL** [165] target long-tail entity coverage. Overall, recent advances focus on improving individual pipeline components rather than proposing wholesale architectural replacements. Within this context, LLMs demonstrate particular value in mitigating **data scarcity** through few-shot knowledge extraction, enabling CH domains previously constrained by limited annotation resources to engage with text-to-KG methodologies more effectively.

> **RQ2: LLMs in text-to-KG pipelines for CH**
>
> How can Large Language Models be integrated into ontology-driven Knowledge Extraction pipelines for Cultural Heritage texts, and what are the limitations, requirements, and trade-offs of such integration?

Chapters 3 and 5 examined the integration of LLMs into text-to-KG pipelines for CH, culminating in the ATR4CH methodology validated in multiple case studies. LLMs can be incorporated as semantic reasoning components within ontology-driven pipelines, complementing symbolic or statistical modules traditionally used for NER and Relation Extraction. Their generative architecture limits character-level span annotation and deterministic output control, but their capacity for contextual reasoning enables robust performance in tasks requiring semantic understanding under conditions of data scarcity. This includes relation classification [210], temporal expression normalization [137], class and type identification for CH ontologies [218], and the handling of implicit information in knowledge extraction [186]. Chapter 5 formalized this integration through the ATR4CH methodology, which coordinates LLM-based extraction with ontological frameworks by transforming three inputs—a corpus of unstructured documents, a target ontology, and a set of Competency Questions—into an operational extraction pipeline refined through iterative evaluation. ATR4CH synthesizes principles from eXtreme Design [158] and the Linked Open Data lifecycle [201], incorporating insights from Odeuropa on ontology-driven annotation design [122] and from MusicBO on evaluation practices [70]. However, this integration reveals several limitations. Because LLMs operate at the token level, they lack deterministic span control and therefore complicate the application of post-processing workflows designed for character-indexed annotation. Their outputs are probabilistic rather than declarative, which hinders standardization and complicates alignment with ontology constraints that demand URI-level precision and hierarchical consistency. In practice, ensuring ontological coherence often requires explicit prompting strategies, schema-aware templates, or hybrid modules performing post-generation validation. These limitations define the practical requirements for successful deployment: prompts must encode domain knowledge and ontological structure, evaluation must be guided by Competency Questions to ensure semantic adequacy, and interdisciplinary collaboration is essential to

align model reasoning with scholarly interpretation. The methodology therefore requires not only computational expertise but also curatorial insight, as the formulation of prompts, examples, and evaluation metrics depends on nuanced understanding of historical and semantic contexts. Within this framework, LLM-based pipelines entail a fundamental trade-off between annotation efficiency and output control. They reduce the need for extensive labeled data and enable few-shot or instruction-based extraction, yet this efficiency comes at the cost of reduced determinism and reproducibility. The resulting balance lies between pipelines favoring automation and flexibility, and those prioritizing strict ontological fidelity at the expense of scalability. ATR4CH positions itself between these extremes by combining LLM-driven extraction with ontology-informed prompt design and structured evaluation, preserving both adaptability and interpretability. In summary, the integration of LLMs into ontology-driven text-to-KG pipelines for CH is both feasible and transformative, provided that their limitations are addressed through methodical design and evaluation. When properly orchestrated, LLMs extend the reach of knowledge extraction into domains where linguistic variability and data scarcity have long constrained semantic enrichment.

> **RQ3: LLMs for scholarly opinion mining**
>
> Can LLM-based extraction systems produce Knowledge Graphs of scholarly interpretations that are sufficiently accurate and complete to answer domain expert competency questions while preserving provenance and epistemic uncertainty?

Chapters 6, 7, and 8 demonstrated that LLM-based systems can produce KGs of adequate quality for domain-specific scholarly applications while revealing systematic performance patterns and limitations requiring human oversight. Chapter 6 implemented ATR4CH over a corpus of 581 Wikipedia articles discussing CH object forgery debates, employing the SEBI ontology to represent authenticity assessment interpretations. The pipeline evaluated

three models (Claude Sonnet 3.7, Llama 3.3 70B, and GPT-4o-mini) across five evaluation dimensions. Models achieved high precision in metadata extraction (F_1 scores between 0.965 and 0.991) and strong evidence extraction quality (accuracy between 95.3% and 96.8%). Entity recognition performance varied substantially (F_1 scores from 0.709 to 0.803), with GPT-4o-mini achieving superior recall (0.912) despite having fewer parameters than Llama 3.3 70B. Hypothesis extraction showed moderate performance (macro F_1 scores between 0.655 and 0.749), with variation reflecting the inherent complexity of scholarly reasoning. G-EVAL scores for discourse representativeness ranged from 0.523 to 0.607. The evaluation revealed critical error propagation: incorrect entity identification consistently compromised downstream extraction accuracy, suggesting that self-consistency checks at the entity recognition stage could substantially improve overall pipeline performance. Chapter 7 extended scholarly opinion extraction to full-length academic articles through a RAG-based approach applied to the *Van den vos Reynaerde* authorship debate. The pipeline processed five scholarly articles, generating nanopublications structured according to the Digital Hermeneutics model with assertion graphs (Layer 1), hermeneutical context graphs (Layer 2), and factual data graphs (Layer 0), yielding 188 factual triples, 49 assertion triples, and 47 methodological triples across the corpus. Competency question coverage analysis demonstrated that three articles addressing authorship attribution produced nanopublications containing sufficient information to answer core research questions about proposed authors, temporal localization, and social contexts, while the two articles which discussed adjacent arguments correctly avoided hallucinated assertions. Domain expert evaluation identified a set of challenges, such as category confusion between work-level entities and manuscript witnesses, temporal uncertainty representation limitations, and occasional false positives caused by relations that were out of scope for the represenation model. The RAG-first architecture successfully addressed document-length bottlenecks by summarizing articles through competency-question-guided retrieval before structured

extraction. Chapter 8 addressed the training data bottleneck through LLM-based generation of semi-synthetic ABSA corpora with DOLCE-aligned entity typing. The methodology produced a dataset of 10,000 book reviews with annotated aspects, categories, and ontological types. Fine-tuning Llama 3.1 8B on this dataset achieved F_1 scores of 0.72 for aspect identification and 0.57 for combined aspect-sentiment detection, with high recall (0.83) indicating effective coverage. Error analysis revealed systematic patterns: strong performance with explicit mentions of agentive elements (characters, authors) but difficulties with implicit references and temporal aspects. Sentiment classification represented the primary bottleneck, particularly for aspects with mixed or nuanced sentiment expressions. The three case studies establish that LLM-based extraction systems can produce KGs sufficiently accurate to support domain expert workflows, though with identifiable limitations. Performance patterns reveal consistent strengths in extracting explicitly stated information while systematic challenges emerge in handling implicit references, nuanced sentiment, and complex reasoning. Provenance preservation proved achievable through reified triple structures linking assertions to source documents and cognizers, though epistemic uncertainty representation requires continued methodological development to capture the full spectrum of scholarly hedging and qualification.

Discussion

The evaluation of LLM-integrated pipelines across multiple case studies reveals performance patterns with practical implications for CH institutions. LLMs achieve accuracy comparable to task-specific tools when deployed for individual subtasks, though they do not surpass state of art components trained on domain-specific annotated data. This performance gap has been consistently documented in the literature: models trained on task-specific data with adequate annotation outperform general-purpose LLMs on those tasks. However, this observation clarifies appropriate deployment contexts rather than invali-

dating LLM integration. CH institutions need not abandon existing annotation practices to benefit from LLMs. LLMs within text-to-KG pipelines provide complementary capabilities across three deployment scenarios: (1) bootstrapping modules during initial development phases or testing architectural decisions before committing annotation resources; (2) augmenting limited training data through synthetic example generation, as demonstrated in Chapter 8; (3) serving as extraction components when the scale of the use case does not justify annotation effort required for supervised training. The third scenario deserves careful consideration. Although LLMs enable extraction without annotation, institutions should prioritize data annotation whenever feasible, even for smaller use cases. Wholesale reliance on LLMs without parallel annotation efforts risks perpetuating data scarcity in CH domains, creating a cycle where lack of training data necessitates LLM use, which in turn reduces motivation for annotation that would enable more accurate task-specific models. Strategic annotation—even at modest scale—builds reusable resources that benefit the broader CH community through shared datasets and enables future development of specialized models as extraction requirements evolve. No single evaluation metric captures all dimensions of text-to-KG quality in CH contexts, as established through validation across multiple case studies. The research employed complementary evaluation strategies addressing different quality dimensions:

- Traditional NLP metrics including precision, recall, and F_1 assess component-level performance, enabling comparison across pipeline stages and identification of extraction bottlenecks;

- Competency Question answering evaluates whether extracted KGs support intended scholarly queries, validating that extraction captures information relevant to research requirements rather than merely achieving high accuracy on arbitrary evaluation sets;

- Back-translation validation detects semantic drift between source text

and extracted representations, as demonstrated in MusicBO [70], using metrics including BLEURT, BLEU, METEOR, BARTScore, and CHRF++ to assess reconstruction fidelity;

- Domain expert validation confirms scholarly utility and identifies systematic errors invisible to automatic metrics, particularly errors involving misrepresentation of evidential relationships or attribution of claims to incorrect sources;
- G-EVAL provides LLM-based discourse representation quality assessment, evaluating whether extracted KGs capture argumentative structures and epistemic uncertainty characteristic of humanities scholarship.

This multi-faceted evaluation addresses heterogeneity identified in RQ1.2, where projects serving different audiences require different validation criteria.

Future Work

This dissertation demonstrated that LLMs can mitigate the data scarcity and linguistic complexity that constrain traditional supervised approaches, while maintaining ontological alignment and scholarly interpretability. Future developments should build on these results along three complementary directions: the creation of reusable infrastructure, the methodological refinement of extraction and evaluation processes, and the extension of ATR4CH to support the representation of scholarly interpretation and epistemic uncertainty. The first direction concerns the transformation of ATR4CH from a methodological framework into a reusable and configurable software environment. The case studies presented in Chapters 4, 6 and 7 demonstrated the adaptability of the methodology across domains; however, each implementation required project-specific development. A configurable framework would allow CH practitioners to specify corpus properties, ontological targets and evaluation criteria declaratively, avoiding the need to implement bespoke extraction logic. Such a system should include modules for document processing across heterogeneous formats, interfaces for interacting with multiple LLM providers

and prompting strategies, management of template-based few-shot examples for domain adaptation, intermediate structured representations for storing extraction results, and evaluation components implementing the multi-metric assessment introduced in Chapter 5. At the architectural level, integration with *SPARQLAnything* [12] would replace the need for ontology-specific serialization code. Instead of developing custom mapping algorithms, LLM outputs in structured JSON could be triplified through the Facade-X meta-model [12], with SPARQL `CONSTRUCT` queries defining the correspondence between extracted content and ontological patterns. This solution would maintain flexibility across different ontologies while simplifying extensibility. By coupling the methodological workflow of ATR4CH with a configuration-driven backend, institutions could reuse established components and focus development efforts on domain modelling and validation rather than low-level RDF generation. A web-based implementation would further lower entry barriers for institutions without dedicated technical staff. Through a graphical interface, practitioners could configure pipeline parameters, upload corpora, inspect extracted triples, and run validation queries interactively. Such an environment would operationalize ATR4CH for real-world use and foster reproducibility of text-to-KG experiments across projects. The second line of development concerns the refinement of methodological components within the ATR4CH pipeline. Standardizing evaluation practices remains a priority: while precision, recall and F_1 scores remain essential for component-level benchmarking, CH applications also require functional assessment through Competency Question answering, back-translation validation, and expert review. A shared repository of benchmark corpora and evaluation templates would promote comparability across projects and support cumulative methodological progress. Further refinement is also needed in the design of LLM prompts and schema alignment strategies. Prompt patterns should be documented alongside annotation guidelines and model configurations, ensuring transparency and reproducibility. Hybrid architectures combining LLM-based extraction with symbolic and/or statis-

tical modules—particularly for entity linking and normalization—may yield more robust pipelines without sacrificing interpretability. These refinements would consolidate ATR4CH as a flexible methodology adaptable to evolving LLM architectures and emerging ontological standards. A third and more exploratory trajectory concerns extending ATR4CH toward the structured representation of scholarly opinions and interpretative claims. The case studies in Chapters 6 and 7 demonstrated the feasibility of using LLMs to extract argumentative structures from humanities discourse, but also revealed persistent difficulties in capturing epistemic nuance. Future work should therefore focus on integrating ontologies capable of representing the interpretative dimension of knowledge. Existing models such as `CRMInf` [51] and `HiCo` [41] provide mechanisms for argumentation and belief adoption, while the recent *Conjectures* model [207] introduces a specialization of Named Graphs distinguishing undisputed, disputed and settled statements. The latter approach enables representation of controversy and provenance within the same graph structure, aligning well with the epistemic requirements of humanities scholarship. Aligning these ontologies with foundational frameworks such as DOLCE or BFO would ensure semantic interoperability while enabling the representation of sourced and unsourced opinions, a limitation of most current sentiment or stance detection models. Developing datasets for scholarly opinion mining will require careful attention to the perspectival nature of interpretation. Annotators may legitimately disagree on opinion boundaries or evidential relationships; rather than treating such disagreement as noise, future datasets should document it explicitly, following perspectivist principles [29]. Combining small-scale manual curation with semi-synthetic augmentation, as demonstrated in Chapter 8, would allow the creation of balanced corpora that preserve interpretative diversity while providing sufficient volume for model adaptation. Fine-tuning open LLMs through parameter-efficient techniques such as LoRA or QLoRA could then support structured generation of Conjectures-compliant KGs, maintaining compatibility with nanopublication

structures for provenance tracking. Evaluation should reflect the plurality of valid interpretations, assessing whether generated graphs preserve argument boundaries, capture legitimate disagreement, and support Competency Questions distinguishing between competing scholarly positions. The long-term objective of these developments is to enable configurable, interpretable, and epistemically-aware text-to-KG systems for the humanities. By transforming ATR4CH into a modular framework, refining its methodological standards, and extending its representational capacity to scholarly interpretation, CH institutions would gain a practical infrastructure for extracting not only factual information but also the interpretative reasoning through which knowledge in the humanities is constructed. Such systems would bridge the gap between large-scale digitization and semantic accessibility, advancing the broader goal of enabling computational hermeneutics grounded in the principles of transparency, provenance, and scholarly validation.

Final Remarks

This dissertation contributes to the fields of Digital Humanities, Semantic Web technologies, and NLP by demonstrating how LLMs can be systematically integrated into ontology-driven text-to-KG pipelines for Cultural Heritage documents. The ATR4CH methodology provides a replicable framework coordinating annotation development, ontological alignment, and LLM-based extraction while maintaining scholarly standards essential for cultural preservation and interpretation as LOD. The research establishes that **LLMs do not replace** traditional text-to-KG pipelines, rather they **augment** them at specific components where data scarcity or task complexity constrain traditional approaches. This finding has practical implications: institutions can adopt LLM-based methods strategically, either augmenting existing pipelines at specific bottlenecks or bootstrapping new implementations for prototyping and testing. Evaluation across multiple case studies demonstrates that LLM-based extraction achieves accuracy sufficient for supporting scholarly queries

while preserving provenance and representing epistemic uncertainty. However, the research also reveals persistent challenges: data (mainly entity linking) and vocabulary alignment, and evaluation. These challenges indicate that human oversight remains necessary, particularly for applications requiring scholarly rigor. The tension identified in the Introduction between digitization's promises and limited semantic accessibility of intellectual content remains partially unresolved. This dissertation advances the state of art by demonstrating feasible approaches to structured knowledge extraction from CH texts. However, systematic solutions enabling institutions to extract, represent, and query humanities knowledge at scale require continued research addressing technical, organizational, and epistemological challenges at the intersection of computational methods and humanities scholarship.

Acknowledgements

Fundings

This research was partially funded by the European Union – Next Generation EU, investment I.4.1 PNRR Patrimonio Culturale, Decreto Ministeriale n. 351, 9 April 2022.

Declaration of AI Use

In the preparation of this dissertation, generative artificial intelligence tools were employed to support text revision, grammar check, content and code creation. Specifically, Claude Sonnet 4.5 (Anthropic, released September 2024) was used for assistance in drafting, structuring, and refining sections of the text.

Personal Acknowledgements

In 2019, one of my last exams for the Bachelor degree in Classical Philology was a laboratory in digital tools for philology. Prof. Francesca Tomasi held a lesson and, while I had some personal familiarity with computers, I struggled to learn everything systematically. It came to a point where passing that exam became a personal challenge. I had to prove to myself that I could understand what was being asked of me. During the exam, I met Arcangelo, who told me about the existence of the Master degree in Digital Humanities. I had never heard of it before, and I was at a crossroads, unsure whether I wanted to continue studying classical philology at the Master's level. Arcangelo explained that the selection process was quite demanding, and once again, the challenge

became personal, as if I had to show someone, perhaps myself, that I could actually do it.

This was essentially the same drive that carried me into this PhD. I barely knew what I was getting into at first. Now I can say, with some disbelief, that I wrote a PhD thesis on it. One thing I have learned over the past three years is that "talent" does not exist in a vacuum. It only takes shape when cultivated daily, within a context. As a kid, I held this belief that if you do not get something right the first time, you are probably not fit for it. That is the most damaging thing you can tell someone, and I cannot help but think of how many real talents have been lost to the mirage of perfectionism.

Almost six years within this department have left me with a deep appreciation for what a community can accomplish when driven by a shared effort. There is too much to say about how much I value the small family we have built. Prof. Tomasi jokes that, even though she started it, it has now taken on a life of its own. One could measure its success by numerous factors: the number of papers published, citation counts, how many international students have joined, how many PhD candidates have come through. But I think the contribution that will stay with us long after the metrics fade is the community that Prof. Tomasi and the rest of us managed to build together. As in every family, we have had our problems too. But if everything ran smoothly and without friction, I do not think we would be as close as we are. Keeping this flame alive and passing it on with each new cycle of students has been both a burden and an honour I shall never forget and I hope to keep doing. Even if the disbelief in some aspects of the systems should warrant more battles.

For this and much more, I owe my deepest gratitude to Prof. Francesca Tomasi. This thesis would not have seen the light without her, the Digital Humanities and Digital Knowledge course within the Department of Classical Philology and Italian Studies and the PhD course of Cultural Heritage in the Digital Ecosystem.

I want to thank Prof. Fabio Vitali for trusting me as a candidate and

for letting me choose my own path, with an always timely steer. Over the course of this journey, I like to think we became not only colleagues but also friends. He is a generous soul who has always worried about and cared for my well-being. It is people like him who make research not just bearable but genuinely enjoyable.

Dr. Marieke van Erp, on the other hand, made sure I did not lose track of what mattered most. Since I was introduced to her during my abroad period, she taught me how to prioritise, how to structure my work with a precise vision, and how to keep moving forward when things felt scattered. The number of times she offered me invaluable advice is uncountable. I think of her as a true mentor, and I cannot imagine how different this path would have been without her presence. She gave the work the sun and water it needed to sprout. I extend this gratitude to the DHLab and the colleagues (Marijn, Joris, Teresa, Andrea...) at the KNAW Humanities Cluster, who welcomed me and made me feel at home during the months I spent in Amsterdam.

I must also thank someone who, while not formally listed as a tutor of this PhD, contributed just as much. Valentina Pasqual started one year before me and, in doing so, guided me closely through each step, experiencing them before most of us others did. She always encouraged me to trust the process and to swallow the bitter pills, because she already knew they would work. She genuinely gives me hope in the future of research in this country, and I know for a fact she will reach any height (literally and metaphorically) she sets her mind to. I am grateful to have her as a colleague and I hope we can keep working together, as both co-authors and friends.

I also want to thank my other colleagues. Lucia Giangolini, with whom I shared this PhD journey and who has been a dear friend since; Stefano, whose wits and fraternal advice I have come to rely on; Arcangelo and Arianna, with whom I started this years ago and I am glad we are still going strong, as even with different paths we keep bouncing our heads back; Teresa, an unexpected friend during my PhD abroad; Ivan, who will never reject helping when needed

just by clicking your heels. There are many others who deserve more thanks than I can fit here. In general, I can only say how grateful I am to have you all in my life. It made things considerably easier and, more importantly, enjoyable.

Valentina Idda, my dearest friend and roommate, who endured my complaints, my doubts, and my occasional spiralling that never made her decide to move out overnight. Her steady presence outside the walls of academia kept me grounded when I needed it most.

I want to thank my parents for their care and genuine interest in my passions. Even though I still sometimes have to explain what exactly it is that I do, I am glad they always listen with curiosity and excitement. They have had my back at every step, and that certainty has meant more than I have probably told them.

Last but not least, I want to thank Lucia for her dear love and appreciation, which has been a constant throughout the ten years we have been together. Even though we live apart for most of the year, I hope we can be together soon. I know for a fact that moment is near. It only truly feels like home when I know I can hear your laugh every morning.

In the end, this life boils down to what we can build together and what we can do to make the next person's burden a little easier to carry. Both within academia, as I strive to do the same Valentina has done for me when I needed it to my younger colleagues, and outside the thick red walls of this beautiful city. Two are better than one, because they have a good reward for their toil. For if they fall, one will lift up his fellow.

The past three years, and likely the ones ahead, have unfolded alongside such sorrow and tragedy around the world that it has, at times, made this work feel entirely trivial. Finding meaning within the borders of a paper discussing how training data should be annotated while bombs, death, and misery thrive outside is something that should drive anyone crazy. This is a memorandum to us all to keep trying to make things better, even if in small steps, and to

do research with that intent in mind. For this, this thesis is dedicated to the children, women, men, and colleagues who were martyred from the river to the sea for no other reason than simply being alive. I can't say much more. See you either in hell or in communism.[13]

— Marco Castello, *Quaglia Sovversiva*, Brano #8, 2:25–2:28

[13]Seriously? Even here? Yeah I know it's a quote from Slavoj Žižek. There you have it.

Appendix

Survey Queries

Primary Methodological Queries

- "knowledge graph extraction" AND "cultural heritage"
- "text to knowledge graph" AND "cultural heritage" OR "humanities"
- "information extraction" AND "knowledge graphs" AND "cultural heritage" OR "humanities"
- "relation extraction" AND "cultural heritage" AND "NLP"
- "entity extraction" AND "digital humanities" AND "RDF"
- "knowledge base population" AND "cultural heritage"
- "semantic extraction" AND "GLAM"
- "ontology alignment" AND "cultural heritage" AND "knowledge extraction"
- "vocabulary alignment" AND "knowledge graphs" AND "humanities" OR "cultural heritage"

Data Alignment Challenges

- "coreference resolution" AND "knowledge graphs" AND "cultural heritage" OR "humanities"

- "entity disambiguation" AND "document" OR "historical texts" AND "knowledge extraction"
- "disambiguation" AND "cultural heritage" AND "knowledge graphs"
- "T-Box alignment" AND "ontology" AND "knowledge extraction"
- "A-Box alignment" AND "RDF" AND "deduplication"
- "URI alignment" AND "entity linking" AND "cultural heritage"
- "CIDOC CRM" AND "ontology mapping" AND "text extraction"
- "schema matching" AND "knowledge graphs" AND "cultural heritage" OR "humanities"

Domain-Specific Challenges

- "historical language" AND "NLP" AND "knowledge extraction"
- "multilingual" AND "knowledge graphs" AND "cultural heritage"
- "long tail entities" AND "cultural heritage" AND "entity linking"
- "terminology" AND "cultural heritage" OR "humanities" AND "entity recognition"
- "temporal knowledge" AND "cultural heritage" AND "extraction"
- "concept drift" AND "knowledge graphs" AND "historical texts"
- "literary texts" AND "knowledge extraction" AND "ambiguity"
- "narrative texts" AND "knowledge graphs" AND "information extraction"
- "TEI" AND "knowledge extraction" AND "semantic annotation"
- "archive" AND "knowledge graphs" AND "extraction"

Document Phenomenology and Text Types

- "literary language"/"cultural heritage" AND "knowledge extraction" AND "computational analysis"
- "metaphor" AND "knowledge graphs" AND "text mining"
- "intertextuality" AND "knowledge extraction" AND "digital humanities" OR "cultural heritage"
- "narrative modeling" AND "knowledge graphs" AND "literary texts"
- "semi-structured documents" AND "knowledge extraction" AND "archives" OR "cultural heritage"
- "interpretative content" AND "knowledge graphs" AND "humanities" OR "cultural heritage"

Methodological Approaches

- "BERT" AND "knowledge extraction" AND "humanities"
- "LLM" AND "knowledge graph construction" AND "cultural heritage"
- "AMR" AND "knowledge graph" AND "text extraction"
- "hybrid approaches" AND "knowledge extraction" AND "humanities"
- "rule-based extraction" AND "cultural heritage" AND "knowledge graphs"
- "frame semantics" AND "knowledge graphs" AND "cultural heritage"
- "provenance tracking" AND "knowledge graphs" AND "digital humanities"
- "human-in-the-loop" AND "knowledge extraction" AND "cultural heritage"

Survey Tables

Table 8.6: BiographySampo: Text-to-KG Implementation

Dimension	BiographySampo Implementation
Project	BiographySampo - Finnish Biographies on the Semantic Web Timeline: 2018–present (launched 09/2018) Domain: Finnish biographical heritage, prosopographical research Scale: 10 million triples
Input	Source: Five biographical collections from Finnish Literature Society Language: Finnish Text Type: Mixed structured biographical summaries with semi-formal notation and free-form narrative text Format: Semi-structured text with pattern-based biographical entries
Methodology	Paradigm: Hybrid approach combining rule-based and NLP techniques Components: Regular expressions for semi-formal sections, Finnish dependency parser, symbolic reasoning module Entity Linking: Name and birth year comparison for external source linking to authority files Graph Construction: Event-based modeling, based on Bio CRM
Representation	Framework: Bio CRM with `:Person` as central class Standards: RDF, CIDOC CRM, Bio CRM, NIF, DCTerms, Schema.org, FOAF, Getty Vocabularies Expressiveness: Event-centric modeling Provenance: ref. to original biographical texts + external databases
Evaluation	Methodology: Manual evaluation of randomly selected sample (135 events from 50 biographies) Metrics: Event identification P=0-99, time A=0.98, place linking P=0.98 and R=0.77 Benchmarking: Accuracy variation by biographical interconnectedness
Output	Format: RDF/Turtle, public SPARQL endpoint Accessibility: Open access semantic portal at biografiasampo.fi Applications: Web applications (faceted search, map visualization, network analysis, statistical tools)
Challenges	Coverage: Accuracy degradation for individuals mentioned across multiple biographies (0.80 vs 0.97%) Quality: OCR errors in digitized sources, lemmatization issues Scalability: Human supervision for complex family relationship extraction
Features	Methodological: pipeline architecture accommodating both structured and unstructured content Integration: Comprehensive external data linking to 16 sources including genealogical, bibliographic, and CH databases Impact: 43,000 users in first five months

Table 8.7: ParliamentSampo: Parliamentary Data as KGs

Dimension	**ParliamentSampo Implementation**
Project	ParliamentSampo - Parliament of Finland on the Semantic Web Timeline: 2020–2023 (launched 02/2023) Domain: Finnish parliamentary archive, political science research Scale: 1 million speeches (1907–2024), 1,800+ members of parliament
Input	Source: Parliament of Finland open data Language: Finnish (primary), Swedish (minority) Text Type: Parliamentary debates, procedural speeches Format: OCR-processed minutes, structured databases, video recordings
Methodology	Paradigm: Hybrid OCR pipeline with NLP processing Components: OCR correction, linguistic annotation, content tagging, speaker identification, entity recognition Entity Linking: Integration with Wikidata, BiographySampo, government databases Graph Construction: Event-based modeling with speeches as central entities
Representation	Framework: Data-driven parliamentary ontology Standards: RDF, Parla-CLARIN, NIF, DCTerms, CIDOC CRM Expressiveness: Temporal modeling, procedural representation, political networks Provenance: Linkage to source documents and media
Evaluation	Methodology: Academic consortium validation, integration testing Metrics: Documented through 20+ peer-reviewed publications Benchmarking: ParlaMint European standards compliance, FIN-CLARIAH integration
Output	Format: RDF/Turtle, Parla-CLARIN XML, public SPARQL endpoint Accessibility: Open portal at parlamenttisampo.fi Applications: Temporal analysis, network visualization, faceted search, statistical tools, geographic analysis
Challenges	Coverage: OCR quality varies for historical documents, automatic annotation (post-2015 speeches) Quality: Mixed speech content correction in 1999 sessions, ongoing OCR error correction Scalability: Query performance optimization for million-speech corpus
Features	Methodological: Data-driven ontology construction from official parliamentary sources Integration: Part of national Sampo portal network and European ParlaMint consortium Temporal Scope: 117-year coverage (1907–2024) with regular updates

Table 8.8: WarSampo: World War II Heritage as Linked Data

Dimension	WarSampo Implementation
Project	WarSampo - Finnish World War II on the Semantic Web Timeline: 2014–present (launched 11/2015) Domain: Finnish World War II military history (1939–1945) Scale: 14 million triples
Input	Source: 20 datasets from National Archives, Defence Forces, National Land Survey, historical societies Language: Finnish, Swedish Text Type: Military records, war diaries, memoirs, official documents, biographical data Format: OCR-processed PDFs, spreadsheets, digital images, XML databases
Methodology	Paradigm: Hybrid approach (OCR processing, rule-based annotation, and semantic integration) Components: linguistic preprocessing, AATOS annotation tool, SPARQL ARPA entity linking, OCR post-processing Entity Linking: KOKO ontology, DBpedia, Wikipedia, specific vocabularies Graph Construction: CIDOC CRM with military-specific extensions
Representation	Framework: Extended CIDOC CRM with domain ontologies for military units, ranks, places, events Standards: RDF, OWL, CIDOC CRM, SPARQL Expressiveness: Military hierarchies, temporal events, geographical relationships, biographical data Provenance: Source attribution to archival documents and media
Evaluation	Methodology: Manual evaluation of 433 magazine articles for NER Metrics: Military unit identification P=0.82, R=0.75with OCR post-processing Benchmarking: Performance comparison before and after OCR correction across entity types
Output	Format: RDF/Turtle, public SPARQL endpoint, CC BY license Accessibility: `https://www.sotasampo.fi/en/` Applications: Storytelling applications with facets, homepage generation for entities
Challenges	Coverage: Historical document digitization quality varies, incomplete records for some military units Quality: OCR errors in historical documents, disambiguation of common names and places Scalability: Processed 26,400 war diaries
Specialty	Methodological: Automatic soldier biography reconstruction from distributed sources Integration: Multi-source data harmonization across 20 institutional datasets

Table 8.9: HyperReal: Cultural Symbolism KG

Dimension	HyperReal Implementation
Project	HyperReal - Marriage is a Peach and a Chalice: Modelling Cultural Symbolism Timeline: 2021 (publication), developed using SAMOD methodology in three iterations Domain: Cultural symbolism across multiple cultures and contexts Scale: 41,416 simulations, 498,525 triples
Input	Source: DBpedia (3,727 simulations), WordNet (81 simulations), Olderr's Symbolism dictionary (37,647 simulations) Language: Multiple languages (primarily English) Text Type: Semi-structured dictionary entries Format: RDF triples from DBpedia, WordNet synsets and definitions
Methodology	Paradigm: Ontology-driven knowledge re-engineering with pattern matching Components: Automatic conversion algorithms, markup-based extraction, NLP modules Entity Linking: Direct mapping to source URIs (DBpedia, WordNet) Graph Construction: Simulation Ontology schema with n-ary relationships, specialized symbolic relationship types
Representation	Framework: Simulation Ontology based on Baudrillard's theory Standards: OWL2-DL, RDF, PROV-O, SKOS Expressiveness: Context-dependent simulations, specialized symbolic relationships, variant relationships Provenance: Full attribution to the encyclopedic entry
Evaluation	Methodology: Competency question testing, automatic ontology evaluation, conversion algorithm validation Metrics: Conversion algorithms achieved F_1=0.97 on 112 manually annotated simulations Benchmarking: FOOPS! ontology evaluation (77% score), HermiT reasoner consistency checking, unit testing framework
Output	Format: OWL2-DL KG, persistent URIs via w3id service Accessibility: `https://www.w3id.org/simulation/data/` Applications: Cultural symbolism research, quantitative analysis of symbolic relationships, cross-cultural symbol comparison
Challenges	Coverage: Copyright restrictions limit public release to subset of data (excludes Olderr's dictionary) Quality: Manual corrections required for source inconsistencies, automatic URI generation errors Scalability: Processing 40,000+ symbolic meanings from heterogeneous sources with varying structures
Specialty	Methodological: First dedicated cultural symbolism KG Theoretical Foundation: Baudrillard's Simulation theory implementation with reusable ontology design patterns Case Study: White roses symbolic analysis demonstrating quantitative research capabilities on cultural symbols

Table 8.10: InTaVia: Transnational Biographical Heritage Integration

Dimension	InTaVia Implementation
Project	InTaVia - In/Tangible European Heritage: Visual Analysis, Curation & Communication Timeline: H2020 project (2021–2024) Domain: European biographical heritage, cultural objects, transnational integration Scale: 24,588,310 triples, 112,050 persons
Input	Source: APIS-Austria, BiographySampo-Finland, SBI-Slovenia, BiographyNet-Netherlands Language: English, Dutch, Finnish, Slovenian Text Type: Biographical dictionaries, Wikipedia articles, CH metadata Format: TEI files, JSON APIs, RDF datasets, external APIs (Europeana, Wikidata)
Methodology	Paradigm: Dual-framework NLP pipeline Components: Flair pipeline (NER, SRL, EL), AllenNLP pipeline (SRL, coreference) Entity Linking: Wikidata, GND, VIAF (federated SPARQL queries) Graph Construction: IDM-RDF ontology, Bio CRM extension, OAI-ORE proxy model
Representation	Framework: IDM-RDF (based on CIDOC CRM v7.1.1, Bio CRM, PROV-O, BIBFRAME, Europeana Data Model) Standards: RDF, OWL, SHACL, OAI-ORE proxy concept for multiple perspectives Expressiveness: Multiple perspectives on entities, contradictory information preservation, temporal uncertainty modeling Provenance: Platform and bibliographic provenance
Evaluation	Methodology: Comparative analysis with DBpedia/Wikidata using *closer reading* Metrics: Captures biographical information not present in existing encyclopedic KGs
Output	Format: RDF/Turtle, JSON API, SPARQL endpoint, Reconciliation Service API Accessibility: intavia.eu SPARQL endpoint, REST API Applications: Data Curation Lab, Visual Analytics Studio, Storytelling Suite
Challenges	Coverage: Limited sameAs links between datasets (548 cross-dataset links), heterogeneous data quality Quality: Unequal serialization across source datasets, missing provenance in some conversions Scalability: problems with computational intensity of enrichment pipelines and federated queries
Specialty	Methodological: Transnational biographical KG harmonization Integration: Multiple perspectives preservation via OAI-ORE proxies, contradictory information modeling

Table 8.11: Odeuropa: Olfactory Heritage Knowledge Extraction

Dimension	Odeuropa Implementation
Project	Odeuropa - Negotiating Olfactory and Sensory Experiences in Cultural Heritage Practice and Research Timeline: H2020 project (2021–2023) Domain: European olfactory heritage, sensory experiences, cultural symbolism Scale: 2.4 million smell instances from 167,029 textual resources + 92,149 instances from 43,679 visual materials
Input	Source: Historical texts and images from European CH collections Language: English, Italian, French, Dutch, German, Slovenian, Latin Text Type: Novels, theatre scripts, travel writing, botanical textbooks, court records, sanitary reports, sermons, medical handbooks Format: Public domain digital texts and images, manually annotated training sets
Methodology	Paradigm: Frame-based annotation with machine learning Components: FrameNet-inspired annotation scheme, multilingual BERT models, computer vision for image analysis Entity Linking: Integration with Flavornet database, Dravnieks descriptors
Representation	Framework: CIDOC CRM and CRMsci with olfactory extensions Standards: RDF, OWL, CIDOC CRM, CRMsci, SPARQL Expressiveness: Frame Elements (Smell Source, Quality, Perceiver, Evoked Odorant, Location, Time, Effect), olfactory gestures, fragrant spaces Provenance: Source attribution to original heritage collections with permalinks
Evaluation	Methodology: Multilingual benchmark annotation across seven European languages Metrics: Performance evaluation of BERT-based models for olfactory information extraction (specific metrics referenced in [133]) Benchmarking: Competency questions from domain experts (70+ questions), validation against olfactory taxonomies
Output	Format: European Olfactory KG, SPARQL endpoint, interactive web interface Accessibility: Smell Explorer at `https://explorer.odeuropa.eu`, downloadable datasets, open source software on GitHub Applications: Three navigation paths (smell sources, fragrant spaces, gestures and allegories), nose-first querying, interactive nosebooks
Challenges	Coverage: Subjective nature of olfactory experiences, historical language variations for smell descriptions Quality: Cross-cultural and cross-temporal consistency in smell terminology, disambiguation of polysemous smell words Scalability: Processing 2.4 million smell instances across multiple languages and modalities
Specialty	Methodological: First comprehensive olfactory heritage KG, multimodal text and image integration Innovation: KG annotation approach with seamless annotation-to-ontology pipelineg Impact: Europa Nostra heritage award winner, novel methodology for sensory heritage digitization

Table 8.12: Notarypedia: Legal Heritage Knowledge Extraction

Dimension	Notarypedia Implementation
Project	Notarypedia - KG Representation of Cultural Heritage Texts Timeline: 2018–2019 (initial prototype), ongoing development Domain: Maltese legal archives, notarial acts, genealogical research Scale: Over 20,000 historical manuscripts (15th century-), 245,901 triples
Input	Source: Notarial Archives in Valletta, transcribed registers from Documentary Sources of Maltese History Language: Latin (primary), medieval Sicilian, Maltese Text Type: Legal documents (wills, property transfers, marriage contracts, receipts) Format: Transcribed historical manuscripts, publication indices
Methodology	Paradigm: Hybrid machine learning and rule-based approach Components: ML models trained on publication indices, rule-based date extraction, document classification Entity Linking: Manual annotation from publication indices, domain-specific keyword identification Graph Construction: Scalable storage without predefined schema, entity-relationship modeling
Representation	Framework: Notarial Ontology incorporating FOAF, Schema.org, Getty Vocabularies Standards: RDF, OWL Expressiveness: Genealogical relationships, geographical connections, commercial relationships Provenance: Source attribution to original manuscripts and publication indices
Evaluation	Methodology: Performance assessment on entity recognition tasks Metrics: Relation extraction, Logistic Regression (F_1=0.68); Entity disambiguation with rule-based approach with F_1=0.96 vs Jaro-Winkler F_1=0.87; TransE for link prediction (A=0.87) Benchmarking: Evaluation against manually annotated publication indices
Output	Format: KG with web interface Accessibility: Planned web interface for KG navigation Applications: Faceted search using keywords for genealogical, historical and legal research
Challenges	Coverage: Multilingual corpus with Latin, medieval Sicilian, and Maltese variants Quality: Incomplete or illegible text in historical manuscripts, scribal inconsistencies Scalability: Processing 20,000+ manuscripts with varying preservation states
Specialty	Methodological: Neurosymbolic approaches using a reasoner for link prediction Integration: Multi-source data enrichment from paleography, conservation Domain Focus: Medieval commercial and legal relationships, trading activity analysis

Table 8.13: MMKG: Musical Meetups KG

Dimension	MMKG Implementation
Project	Musical Meetups KG (MMKG) - Historical Social Network Analysis Timeline: 2022–2025 (Polifonia project) Domain: European musical heritage, historical social networks, biographical analysis Scale: 45,812 historical meetups from 33,309 biographies
Input	Source: Wikipedia biographies of musical artists (`dbo:MusicalArtist` entities) Language: Primarily English (Wikipedia corpus) Text Type: Biographical narratives from European musical culture (1800–1945) Format: Text-based Wikipedia articles with sentence-level indexing and paragraph organization
Methodology	Paradigm: Hybrid pipeline (NLP, knowledge engineering, machine learning, LLMs) Components: DBpedia Spotlight, SynTime-based temporal extraction, GPT for temporal normalization, SpaCy coreferee Entity Linking: DBpedia resource linking, quality control filtering Graph Construction: MEETUPS ontology with harmonization algorithm for consecutive sentence processing
Representation	Framework: MEETUPS ontology incorporating Time Ontology, PROV, SEM, Polifonia CORE Standards: RDF, SPARQL Expressiveness: Historical meetups (participants, locations, time, purpose) Provenance: Source attribution to original biographical sentences
Evaluation	Methodology: Competency question testing, domain expert survey (12 participants), case study analysis Metrics: Purpose classification (P=0.85, LLM-enhanced vs 0.46 ML-only) Benchmarking: Expert validation showing 100% agreement on encounter documentation importance, 75% acknowledge KG teaching utility
Output	Format: RDF (N-quads), SPARQL endpoint Accessibility: SPARQL endpoint at polifonia.kmi.open.ac.uk Applications: Historical social network analysis, exploratory data analysis, educational tools, GIS integration, timeline visualization
Challenges	Coverage: Entity ambiguity in DBpedia Spotlight, temporal expression complexity, implicit coreference detection Quality: Name variations, entity disambiguation, 35% temporal expressions requiring LLM processing Scalability: Processing 33,000+ biographies, harmonization algorithm complexity
Specialty	Methodological: First comprehensive musical encounter KG, harmonization algorithm for multi-sentence meetups Innovation: LLM-enhanced temporal processing, zero-shot purpose classification Impact: Filled resource gap in music history, enables macro-scale cultural exchange analysis

Table 8.14: MusicBO: Bologna Musical Heritage KG

Dimension	MusicBO Implementation
Project	MusicBO - Musical Heritage KG for Bologna Timeline: 2022–2025 (Polifonia project), active development Domain: Bologna musical heritage, diachronic musical documentation (1700s–present) Scale: 531,073 triples from 137 documents
Input	Source: Corpus of musical heritage documents (correspondence, biographies, musicological studies) Language: English, Italian, French, Spanish) Text Type: Historical documents (18th century-present) Format: PDF, images, DOCX files
Methodology	Paradigm: Text2AMR2FRED pipeline with back-translation evaluation Components: SPRING (English) and USeA (Italian) for AMR parsing, AMR2FRED for RDF conversion, BLINK entity linking Entity Linking: BLINK to DBpedia and Wikidata Graph Construction: Event-centric AMR graphs converted to RDF/OWL with named graph tracking to source sentences
Representation	Framework: FRED theoretical framework with PropBank Frames, enriched through Framester alignments Standards: RDF/OWL, named graphs, PropBank predicate-argument structures Expressiveness: Event-centric semantic representation capturing musical heritage relationships and activities Provenance: Named graphs-based tracking to source sentence in corpus
Evaluation	Methodology: Back-translation validation using BLEURT (English) and cosine similarity (Italian) Metrics: Quality filtering retained 7,557 of 62,377 pairs (12.1% retention rate, BLEURT threshold >0, cosine similarity threshold >0.90) Benchmarking: Automated quality assessment for historical text processing
Output	Format: RDF/OWL KG, SPARQL endpoint Accessibility: Public SPARQL endpoint at polifonia.disi.unibo.it/musicbo/sparql Applications: Large-scale qualitative analysis, visual data stories through MELODY, scholarly research on Bologna musical heritage
Challenges	Coverage: NIL and Long-tail entities, domain-specific terminology, historical language variations Quality: OCR errors in historical documents, AMR parsing accuracy for non-standard texts, high filtering requirements Scalability: Intensive quality control filtering (87.9% rejection rate), computational complexity of AMR processing
Specialty	Methodological: Back-translation validation for historical text processing Innovation: Automated quality control for AMR graphs Integration: MELODY visualization platform

Table 8.15: Viewsari: Renaissance Art History KG

Dimension	Viewsari Implementation
Project	Viewsari - Renaissance Social Networks from Vasari's Lives of the Artists Timeline: 2024, ongoing development and analysis Domain: Renaissance art history, biographical networks, cultural heritage documentation Scale: Complete analysis of Vasari's 10-volume work, social networks across XIIIth to XVIth century artists
Input	Source: Giorgio Vasari's "Lives of the Most Eminent Painters, Sculptors, and Architects" (English translation by Gaston C. Du Vere, 1912) Language: English Text Type: Historical biographical narratives, art-historical documentation, literary work Format: Digitized text with Index of Names, paragraph-segmented content with page numbering
Methodology	Paradigm: Hybrid approach combining NER, coreference resolution, and social network analysis Components: NER with LUKE, F-Coref coreference resolution, Index of Names entity resolution, PMI calculation Entity Linking: to VIAF and Wikidata, ICONCLASS motif classification Graph Construction: Three-layered FRBR-inspired ontology with Ontology Design Patterns, co-occurrence-based relationship extraction
Representation	Framework: FRBR-inspired three-layer model (Work, Instantiation, Content levels) with eXtreme Design methodology Standards: RDF, OWL, Linked Open Data vocabularies, SPARQL Expressiveness: First-class co-occurrence entities, provenance-content patterns, social network structures, centrality measures Provenance: Provenance-Content Pattern implementation
Evaluation	Methodology: Network analysis validation, PMI-based relationship strength assessment, centrality analysis Metrics: High PMI values correlate with documented artistic collaborations, betweenness centrality confirms art-historical importance Benchmarking: Cross-validation with established art-historical knowledge, authority file reconciliation
Output	Format: KG (RDF/OWL), interactive network visualizations, SPARQL endpoint Accessibility: Online visualization platform, downloadable datasets via Zenodo, KG with SPARQL queries Applications: Art-historical network analysis, distant reading
Challenges	Coverage: Long-tail entity linking for Renaissance artists and locations, historical name variations Quality: Co-occurrence interpretation limitations, temporal relationship modeling, authority file integration Scalability: Multi-volume text processing, comprehensive Index of Names utilization
Specialty	Methodological: First-class co-occurrence modeling, PMI-based relationship quantification, comprehensive Renaissance network extraction Innovation: Index of Names exploitation for entity resolution Impact: Quantitative validation of art-historical networks, computational discovery of influence patterns, distant reading methodology for art history

Table 8.16: Bernoulli-Euler Digital: Historical Scientific Correspondence Knowledge Extraction

Dimension	Bernoulli-Euler Digital Implementation
Project	Bernoulli-Euler Digital - RDF-based Digital Editions for Historical Scientific Correspondence Timeline: Built upon BEOL (2016–2020), enhanced 2022–2023 with text-to-KG pipeline Domain: Early modern mathematics and science, scientific correspondence, travel documentation (18th–19th centuries) Scale: 1,946 letters, 8 million RDF triples
Input	Source: Multiple historical editions including BEBB, LEOO IVA/IV, LECE, Jacob Bernoulli's *Meditationes* Language: Primarily German, French, Latin with some English translations Text Type: Scientific correspondence, travel diaries Format: TEI/XML, mathematical notation
Methodology	Paradigm: Hybrid approach (NER, supervised RE, ontological constraints) Components: textToRDFGraph pipeline, GND/Geonames entity linking, RDF-star for provenance Entity Linking: Integration with GND, Geonames, Wikidata Graph Construction: modeling with CIDOC CRM extensions, specialized classes for scientific correspondence
Representation	Framework: Extended CIDOC CRM, BIBLIO for publications Standards: RDF-star for statement-level metadata, IIIF, TEI, SPARQL Expressiveness: LaTeX notation, geographical relationships, epistolary networks, bibliographical references, temporal modeling Provenance: RDF-star triples linking extracted knowledge to source documents, standoff markup preserving text positions
Evaluation	Methodology: Integration testing with research infrastructure, validation through scholarly use cases Metrics: P=0.99 for event identification, A=0.98 accuracy for temporal information, network visualization validation Benchmarking: Cross-validation with established historical sources, integration with international correspondence databases (EMLO)
Output	Format: RDF/Turtle with RDF-star, TEI/XML, SPARQL endpoint Accessibility: Public platform at bernoulli-euler.dhlab.unibas.ch, ARK persistent identifiers Applications: Interactive 3D network visualization, sophisticated query capabilities, SDE platform
Challenges	Coverage: Multilingual corpus with historical language variations, mathematical notations Quality: OCR quality variation in historical documents, entity disambiguation across correspondence Scalability: Processing 8 mln triples with standoff markup, performance on sophisticated queries
Specialty	Methodological: RDF-star implementation for provenance, standoff markup for annotations Integration: Unified platform, connection to broader correspondence networks Innovation: Mathematical notation preservation in triples, 3D force-directed graph visualizations, travel pattern extraction

Bibliography

[1] ACM. *ACM Digital Library*. https://dl.acm.org/. Accessed: 2025-09-18.

[2] ADHO. *Alliance of Digital Humanities Organizations*. https://adho.org/. Accessed: 2025-09-18.

[3] AIUCD. *AIUCD Mailing List*. https://www.aiucd.it/. Accessed: 2025-09-18.

[4] AIUCD. *Associazione per l'Informatica Umanistica e la Cultura Digitale*. https://www.aiucd.it/. Accessed: 2025-09-18.

[5] Akbik, Alan, Blythe, Duncan, and Vollgraf, Roland. "Contextual String Embeddings for Sequence Labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Ed. by Bender, Emily M., Derczynski, Leon, and Isabelle, Pierre. Association for Computational Linguistics, 2018, pp. 1638–1649. URL: https://aclanthology.org/C18-1139/.

[6] Alassi, Sepideh. *From unstructured texts to RDF-star-based open research data queryable by references*. Publisher: Zenodo. 2023. DOI: 10.5281/zenodo.8107643. URL: https://zenodo.org/records/8107643 (visited on 09/20/2024).

[7] Alassi, Sepideh and Rosenthaler, Lukas. " Semantic precision: crafting RDF-based digital editions for unveiling the layers of historical correspondence ". In: *Digital Scholarship in the Humanities* 39.3 (2024), pp. 813–835. ISSN: 2055-7671. DOI: 10.1093/llc/fqae027. eprint:

https://academic.oup.com/dsh/article-pdf/39/3/813/58997802/fqae027.pdf. URL: https://doi.org/10.1093/llc/fqae027.

[8] Alfaifi, Yousef. "Ontology Development Methodology: A Systematic Review and Case Study". In: *2022 2nd International Conference on Computing and Information Technology (ICCIT)*. 2022, pp. 446–450. DOI: 10.1109/ICCIT52419.2022.9711664.

[9] Ali, Manzoor, Saleem, Muhammad, Moussallem, Diego, Sherif, Mohamed Ahmed, and Ngonga Ngomo, Axel-Cyrille. "RELD: A Knowledge Graph of Relation Extraction Datasets". In: *The Semantic Web*. Ed. by Pesquita, Catia, Jimenez-Ruiz, Ernesto, McCusker, Jamie, Faria, Daniel, Dragoni, Mauro, Dimou, Anastasia, Troncy, Raphael, and Hertling, Sven. Cham: Springer Nature Switzerland, 2023, pp. 337–353. ISBN: 978-3-031-33455-9.

[10] Álvarez López, Tamara, Fernández Gavilanes, Milagros, Costa Montenegro, Enrique, Juncal Martínez, Jonathan, García Méndez, Silvia, Bellot, Patrice, et al. "A book reviews dataset for Aspect-Based Sentiment Analysis". In: *Language & Technology Conference, Poznań, Polonia, 17-19 noviembre 2017*. Enxeñaría telemática. 2017.

[11] Andrews, T. "The Structured Assertion Record (STAR) Model for Event-based Representation of Historical Information". In: *GrapHNR 2023*. 2023. URL: https://graphentechnologien.hypotheses.org/files/2023/05/GrapHNR-2023-32-Andrews-STAR.pdf.

[12] Asprino, Luigi, Daga, Enrico, Gangemi, Aldo, and Mulholland, Paul. "Knowledge Graph Construction with a Façade: A Unified Method to Access Heterogeneous Data Sources on the Web". In: *ACM Trans. Internet Technol.* 23.1 (2023). ISSN: 1533-5399. DOI: 10.1145/3555312. URL: https://doi.org/10.1145/3555312.

[13] Baas, Jurian. "Entity Resolution on Historical Knowledge Graphs". English. Doctoral thesis 1 (Research UU / Graduation UU). Universiteit Utrecht, 2023. ISBN: 978-94-6416-395-7. DOI: 10.33540/1933.

[14] Banarescu, Laura, Bonial, Claire, Cai, Shu, Georgescu, Madalina, Griffitt, Kira, Hermjakob, Ulf, Knight, Kevin, Koehn, Philipp, Palmer, Martha, and Schneider, Nathan. "Abstract Meaning Representation for Sembanking". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Ed. by Pareja-Lora, Antonio, Liakata, Maria, and Dipper, Stefanie. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 178–186. URL: https://aclanthology.org/W13-2322/.

[15] Banerjee, Satanjeev and Lavie, Alon. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by Goldstein, Jade, Lavie, Alon, Lin, Chin-Yew, and Voss, Clare. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[16] Barabucci, G., Tomasi, F., and Vitali, F. "Supporting complexity and conjectures in cultural heritage descriptions". In: *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*. CEUR Workshop, 2021, pp. 104–115. URL: http://ceur-ws.org/Vol-2810/.

[17] Baroncini, S., Sartini, B., Erp, M. van, Tomasi, F., and Gangemi, A. "Is dc:subject enough? A landscape on iconography and iconology statements of knowledge graphs in the semantic web". In: *Journal of Documentation* 79 (2023), pp. 115–136. DOI: 10.1108/JD-09-2022-0207.

[18] Barone, N. "Intorno alla falsificazione dei documenti ed alla critica di essi. Memoria letta all'Accademia Pontaniana nella tornata del 21 gen-

naio 1912". In: *Atti Dell'Accademia Pontaniana* 42 (1912). URL: http://www.rmoa.unina.it/4359/.

[19] Ben Abacha, Asma, Yim, Wen-wai, Fu, Yujuan, Sun, Zhaoyi, Yetisgen, Meliha, Xia, Fei, and Lin, Thomas. "MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Che, Wanxiang, Nabende, Joyce, Shutova, Ekaterina, and Pilehvar, Mohammad Taher. Vienna, Austria: Association for Computational Linguistics, 2025, pp. 22539–22550. ISBN: 979-8-89176-256-5. DOI: 10.18653/v1/2025.findings-acl.1159. URL: https://aclanthology.org/2025.findings-acl.1159/.

[20] Bender, Emily M. and Friedman, Batya. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science". In: *Transactions of the Association for Computational Linguistics* 6 (2018). Ed. by Lee, Lillian, Johnson, Mark, Toutanova, Kristina, and Roark, Brian, pp. 587–604. DOI: 10.1162/tacl_a_00041. URL: https://aclanthology.org/Q18-1041/.

[21] Berardinis, Jacopo de, Carriero, Valentina Anita, Jain, Nitisha, Lazzari, Nicolas, Meroño-Peñuela, Albert, Poltronieri, Andrea, and Presutti, Valentina. "The Polifonia Ontology Network: Building a Semantic Backbone for Musical Heritage". In: *The Semantic Web – ISWC 2023*. Vol. 14266. Lecture Notes in Computer Science. Springer, 2023, pp. 302–322. DOI: 10.1007/978-3-031-47243-5_17. URL: https://doi.org/10.1007/978-3-031-47243-5_17.

[22] Berners-Lee, Tim, Hendler, James, and Lassila, Ora. "The semantic web". In: *Scientific American* 284.5 (2001), pp. 34–43. DOI: 10.1038/scientificamerican0501-34.

[23] Besamusca, Bart and Bouwman, André. *Of Reynaert the Fox: Text and Facing Translation of the Middle Dutch Beast Epic Van den vos Rey-*

naerde. Open access. Amsterdam: Amsterdam University Press, 2009, p. 368. ISBN: 9789089640246. DOI: 10.5117/9789089640246. URL: http://library.oapen.org/handle/20.500.12657/35321.

[24] Blau, N. "Uncertainty and the history of ideas". In: *History and Theory* 50.3 (2011), pp. 358–372. DOI: 10.1111/j.1468-2303.2011.00590.x.

[25] Boer, Victor de. "Knowledge Graphs for Cultural Heritage and Digital Humanities". In: *Proceedings of the 5th Workshop on AnalySis, Understanding and ProMotion of HeritAge Contents*. SUMAC '23. Ottawa ON, Canada: Association for Computing Machinery, 2023, p. 3. ISBN: 9798400702792. DOI: 10.1145/3607542.3617354. URL: https://doi.org/10.1145/3607542.3617354.

[26] Borgo, Stefano, Ferrario, Roberta, Gangemi, Aldo, Guarino, Nicola, Masolo, Claudio, Porello, Daniele, Sanfilippo, Emilio M., Vieu, Laure, Borgo, Stefano, Galton, Antony, and Kutz, Oliver. "DOLCE: A descriptive ontology for linguistic and cognitive engineering1". In: *Appl. Ontol.* 17.1 (2022), pp. 45–69. ISSN: 1570-5838. DOI: 10.3233/AO-210259. URL: https://doi.org/10.3233/AO-210259.

[27] Brown, Tom B., Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, Agarwal, Sandhini, Herbert-Voss, Ariel, Krueger, Gretchen, Henighan, Tom, Child, Rewon, Ramesh, Aditya, Ziegler, Daniel M., Wu, Jeffrey, Winter, Clemens, Hesse, Christopher, Chen, Mark, Sigler, Eric, Litwin, Mateusz, Gray, Scott, Chess, Benjamin, Clark, Jack, Berner, Christopher, McCandlish, Sam, Radford, Alec, Sutskever, Ilya, and Amodei, Dario. "Language models are few-shot learners". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[28] Busa, Roberto. "The Annals of Humanities Computing: The Index Thomisticus". In: *Computers and the Humanities* 14.2 (1980), pp. 83–90. URL: http://www.jstor.org/stable/30207304.

[29] Cabitza, Federico, Campagner, Andrea, and Basile, Valerio. "Toward a perspectivist turn in ground truthing for predictive computing". In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN: 978-1-57735-880-0. DOI: 10.1609/aaai.v37i6.25840. URL: https://doi.org/10.1609/aaai.v37i6.25840.

[30] Carletta, Jean. "Assessing Agreement on Classification Tasks: The Kappa Statistic". In: *Computational Linguistics* 22.2 (1996). Ed. by Hirschberg, Julia, pp. 249–254. URL: https://aclanthology.org/J96-2004/.

[31] Carlson, Andrew, Betteridge, Justin, Kisiel, Bryan, Settles, Burr, Hruschka Jr., Estevam R., and Mitchell, Tom M. "Toward an Architecture for Never-Ending Language Learning". In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI'10. 2010, pp. 1306–1313. URL: https://dl.acm.org/doi/10.5555/2898607.2898816.

[32] Carriero, Valentina Anita, Mariani, Fabio, Nuzzolese, Andrea Giovanni, Pasqual, Valentina, and Presutti, Valentina. "Agile Knowledge Graph Testing with TESTaLOD". In: *ISWC (Satellites)*. 2019, pp. 221–224. URL: https://ceur-ws.org/Vol-2456/paper58.pdf.

[33] Carroll, J.J., Bizer, C., Hayes, P., and Stickler, P. "Named graphs". In: *Journal of Web Semantics* 3.4 (2005), pp. 247–267. DOI: 10.1016/j.websem.2005.09.001.

[34] Chebolu, Siva Uday Sampreeth, Dernoncourt, Franck, Lipka, Nedim, and Solorio, Thamar. "A Review of Datasets for Aspect-based Sentiment Analysis". In: *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Park, Jong C., Arase, Yuki, Hu, Baotian, Lu, Wei, Wijaya, Derry, Purwarianti, Ayu, and Krisnadhi, Adila Alfa. Nusa Dua, Bali: Association for Computational Linguistics, 2023, pp. 611–628. DOI: 10.18653/v1/2023.ijcnlp-main.41. URL: https://aclanthology.org/2023.ijcnlp-main.41/.

[35] Chia, Yew Ken, Bing, Lidong, Aljunied, Sharifah Mahani, Si, Luo, and Poria, Soujanya. "A Dataset for Hyper-Relational Extraction and a Cube-Filling Approach". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022. URL: https://arxiv.org/abs/2211.10018.

[36] Ciatto, Giovanni, Agiollo, Andrea, Magnini, Matteo, and Omicini, Andrea. "Large Language Models as Oracles for Instantiating Ontologies with Domain-Specific Knowledge". In: *Knowledge-Based Systems* 310 (2025), p. 112940. DOI: 10.1016/j.knosys.2024.112940. URL: https://doi.org/10.1016/j.knosys.2024.112940.

[37] Cigliano, Andrea and Fallucchi, Francesca. "The Convergence of Open Data, Linked Data, Ontologies, and Large Language Models: Enabling Next-Generation Knowledge Systems". In: *Metadata and Semantic Research*. Ed. by Sfakakis, Michalis, Garoufallou, Emmanouel, Damigos, Matthew, Salaba, Athena, and Papatheodorou, Christos. Springer Nature Switzerland, 2025, pp. 197–213. ISBN: 978-3-031-81974-2. DOI: https://doi.org/10.1007/978-3-031-81974-2_17.

[38] Ciotti, Fabio, Tomasi, Francesca, Daquino, Marilena, and Lana, Maurizio. "Using ontologies as a faceted browsing for heterogeneous cultural

heritage collections". In: *CEUR Workshop Proceedings*. CEUR-WS.org, 2015. URL: https://hdl.handle.net/11579/72102.

[39] Cohen, Jacob. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104. eprint: https://doi.org/10.1177/001316446002000104. URL: https://doi.org/10.1177/001316446002000104.

[40] Daquino, Marilena, Pasqual, Valentina, and Tomasi, Francesca. "Knowledge Representation of digital Hermeneutics of archival and literary Sources". In: *JLIS.it* 11.3 (2020), pp. 59–76. DOI: 10.4403/jlis.it-12642. URL: https://www.jlis.it/index.php/jlis/article/view/35.

[41] Daquino, Marilena and Tomasi, Francesca. "Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects". In: *Metadata and Semantics Research. MTSR 2015*. Ed. by Garoufallou, E., Hartley, R., and Gaitanou, P. Vol. 544. Communications in Computer and Information Science. Springer, 2015. DOI: 10.1007/978-3-319-24129-6_37.

[42] DBLP. *DBLP Computer Science Bibliography*. https://dblp.org/. Accessed: 2025-09-18.

[43] Devedzic, Vladan. "Education and the Semantic Web". In: *International Journal of Artificial Intelligence in Education* 14.2 (2004), pp. 165–191. DOI: 10.3233/IRG-2004-14(2)02. eprint: https://journals.sagepub.com/doi/pdf/10.3233/IRG-2004-14%282%2902. URL: https://journals.sagepub.com/doi/abs/10.3233/IRG-2004-14%5C%282%5C%2902.

[44] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". 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). Ed. by Burstein, Jill, Doran, Christy, and Solorio, Thamar. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[45] *Digital Humanities Atlas*. https://dh-atlas.github.io/. Accessed: 2025-09-18.

[46] *Digital Humanities Benelux*. https://dhbenelux.org/. Accessed: 2025-09-18.

[47] Di Pasquale, Alessio, Pasqual, Valentina, Tomasi, Francesca, and Vitali, Fabio. "On assessing weaker logical status claims in Wikidata cultural heritage records". In: *Semantic Web Journal* (2024). URL: https://hdl.handle.net/11585/1004003.

[48] Diaz-Garcia, Jose A. and Diaz Lopez, Julio Amador. "A survey on cutting-edge relation extraction techniques based on language models". In: *Artificial Intelligence Review* 58.287 (2025). DOI: 10.1007/s10462-025-11280-0. URL: https://doi.org/10.1007/s10462-025-11280-0.

[49] *Catalogue of Digital Editions*. https://dig-ed-cat.acdh.oeaw.ac.at/. Accessed: 2025-09-18.

[50] Dimou, Anastasia, Vander Sande, Miel, Colpaert, Pieter, Verborgh, Ruben, Mannens, Erik, and Van de Walle, Rik. "RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data". In: *Proceedings of the 7th Workshop on Linked Data on the Web*. Ed. by Bizer, Christian, Heath, Tom, Auer, Sören, and Berners-Lee, Tim. Vol. 1184. CEUR Workshop Proceedings. 2014. URL: http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

[51] Doerr, Martin. "The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata". In: *AI magazine.* 24.3 (2003). ISSN: 0738-4602.

[52] Dong, Li, Wei, Furu, Tan, Chuanqi, Tang, Duyu, Zhou, Ming, and Xu, Ke. "Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Association for Computational Linguistics, 2014, pp. 49–54. URL: https://aclanthology.org/P14-2009.

[53] Doumanas, Dimitrios, Bouchouras, Georgios, Soularidis, Andreas, Kotis, Konstantinos, and Vouros, George. "From Human- to LLM-Centered Collaborative Ontology Engineering". In: *Applied Ontology* 19.4 (2024), pp. 334–367. DOI: 10.1177/15705838241305067. URL: https://doi.org/10.1177/15705838241305067.

[54] Douze, Matthijs, Guzhva, Alexandr, Deng, Chengqi, Johnson, Jeff, Szilvasy, Gergely, Mazaré, Pierre-Emmanuel, Lomeli, Maria, Hosseini, Lucas, and Jégou, Hervé. "THE FAISS LIBRARY". In: *IEEE Transactions on Big Data* (2025), pp. 1–17. DOI: 10.1109/TBDATA.2025.3618474.

[55] Drucker, Johanna. "Humanistic Theory and Digital Scholarship". In: *Debates in the Digital Humanities.* Ed. by Gold, Matthew K. Minneapolis, MN: University of Minnesota Press, 2012, pp. 85–95. ISBN: 9780816677948. DOI: 10.5749/minnesota/9780816677948.003.0011.

[56] Dubey, Abhimanyu, Jauhri, Abhinav, and Pandey, Abhinav. *The Llama 3 Herd of Models.* 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[57] Ehrmann, Maud, Hamdi, Ahmed, Pontes, Elvys Linhares, Romanello, Matteo, and Doucet, Antoine. "Named Entity Recognition and Classification in Historical Documents: A Survey". In: *ACM Computing Sur-*

veys 56 (2021), pp. 1–47. URL: https://api.semanticscholar.org/CorpusID:237605282.

[58] *International Conference on Knowledge Engineering and Knowledge Management*. https://ekaw2024.inf.unibz.it/. Accessed: 2025-09-18.

[59] Elango, Pradheep. *Coreference Resolution: A Survey*. Tech. rep. 1.12. Madison, WI: University of Wisconsin, 2005, p. 12.

[60] Ellul, Charlene, Azzopardi, Joel, and Abela, Charlie. "NotaryPedia: A Knowledge Graph of Historical Notarial Manuscripts". In: *On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings*. Rhodes, Greece: Springer-Verlag, 2019, pp. 626–645. ISBN: 978-3-030-33245-7. DOI: 10.1007/978-3-030-33246-4_39. URL: https://doi.org/10.1007/978-3-030-33246-4_39.

[61] Elsevier. *Journal of Web Semantics*. https://www.journals.elsevier.com/journal-of-web-semantics. Accessed: 2025-09-18.

[62] Elsevier. *ScienceDirect*. https://www.sciencedirect.com/. Accessed: 2025-09-18.

[63] *Extended Semantic Web Conference*. https://2024.eswc-conferences.org/. Accessed: 2025-09-18.

[64] Fader, Anthony, Soderland, Stephen, and Etzioni, Oren. "Identifying Relations for Open Information Extraction". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Ed. by Barzilay, Regina and Johnson, Mark. Association for Computational Linguistics, 2011, pp. 1535–1545. URL: https://aclanthology.org/D11-1142/.

[65] Fan, Wenqi, Ding, Yujuan, Ning, Liangbo, Wang, Shijie, Li, Hengyun, Yin, Dawei, Chua, Tat-Seng, and Li, Qing. "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '24. Barcelona, Spain: Association for Computing Machinery, 2024, pp. 6491–6501. ISBN: 9798400704901. DOI: 10.1145/3637528.3671470. URL: https://doi.org/10.1145/3637528.3671470.

[66] Fernández-López, Mariano, Gómez-Pérez, Asunción, and Juristo, Natalia. "METHONTOLOGY: From Ontological Art Towards Ontological Engineering". In: *AAAI Conference on Artificial Intelligence*. 1997. URL: https://api.semanticscholar.org/CorpusID:10550105.

[67] Frey, Johannes, Meyer, Lars, Arndt, Natanael, Brei, Felix, and Bulert, Kirill. "Benchmarking the Abilities of Large Language Models for RDF Knowledge Graph Creation and Comprehension: How Well Do LLMs Speak Turtle?" In: *ArXiv* abs/2309.17122 (2023).

[68] Gadamer, Hans G. *Truth and Method*. Bloomsbury Publishing, 2013. ISBN: 9781780936000.

[69] Gangemi, Aldo. "Ontology Design Patterns for Semantic Web Content". In: *The Semantic Web – ISWC 2005: 4th International Semantic Web Conference, Galway, Ireland, November 6-10, 2005, Proceedings*. Lecture Notes in Computer Science. Springer. 2005, pp. 262–276. DOI: 10.1007/11574620_21.

[70] Gangemi, Aldo, Graciotti, Arianna, Marzi, Eleonora, Meloni, Antonello, Nuzzolese, Andrea, Presutti, Valentina, Reforgiato Recupero, Diego, Russo, Alessandro, and Tripodi, Rocco. " MusicBO, an application of Text2AMR2FRED to the Musical Heritage domain ". In: *20th Extended Semantic Web Conference*. CEUR Workshop Proceedings, 2024. URL: https://doi.org/10.1007/978-3-031-78952-6_29.

[71] Gangemi, Aldo, Graciotti, Arianna, Meloni, Antonello, Nuzzolese, Andrea Giovanni, Presutti, V., Recupero, D., Russo, Alessandro, and Tripodi, Rocco. "Text2AMR2FRED, a Tool for Transforming Text into RDF/OWL Knowledge Graphs via Abstract Meaning Representation". In: *CEUR Workshop Proceedings*. 2023. URL: https://www.semanticscholar.org/paper/Text2AMR2FRED%2C-a-Tool-for-Transforming-Text-into-Gangemi-Graciotti/129ae0493f3702d868bd9106a594deeca0d08724 (visited on 09/17/2024).

[72] Gao, Z., Feng, A., Song, X., and Wu, X. "Target-Dependent Sentiment Classification With BERT". In: *IEEE Access* 7 (2019), pp. 154290–154299. DOI: 10.1109/ACCESS.2019.2946594.

[73] Gardent, Claire, Shimorina, Anastasia, Narayan, Shashi, and Perez-Beltrachini, Laura. "Creating Training Corpora for NLG Micro-Planners". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017, pp. 179–188. DOI: 10.18653/v1/P17-1017. URL: https://www.aclweb.org/anthology/P17-1017.pdf.

[74] Gardner, Matt, Grus, Joel, Neumann, Mark, Tafjord, Oyvind, Dasigi, Pradeep, Liu, Nelson F., Peters, Matthew, Schmitz, Michael, and Zettlemoyer, Luke. "AllenNLP: A Deep Semantic Natural Language Processing Platform". In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Ed. by Park, Eunjeong L., Hagiwara, Masato, Milajevs, Dmitrijs, and Tan, Liling. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 1–6. DOI: 10.18653/v1/W18-2501. URL: https://aclanthology.org/W18-2501/.

[75] Garijo, Daniel, Poveda-Villalón, María, Amador-Domínguez, Elvira, Wang, Ziyuan, García-Castro, Raúl, and Corcho, Oscar. "LLMs for Ontology Engineering: A Landscape of Tasks and Benchmarking Challenges". In: *Proceedings of the 23rd International Semantic Web Con-*

ference (ISWC 2024). Special Session on LLMs. Baltimore, MD, USA: CEUR-WS.org, 2024.

[76] Gerstgrasser, Matthias, Schaeffer, Rylan, Dey, Apratim, Rafailov, Rafael, Sleight, Henry, Hughes, John, Korbak, Tomasz, Agrawal, Rajashree, Pai, Dhruv, Gromov, Andrey, Roberts, Daniel A., Yang, Diyi, Donoho, David L., and Koyejo, Oluwasanmi. "Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data". In: *ArXiv* abs/2404.01413 (2024). URL: https://api.semanticscholar.org/CorpusID:268856698.

[77] Giagnolini, Lucia, Bonora, Paolo, and Tomasi, Francesca. "Affinare il contesto: estrazione di informazioni strutturate per l'arricchimento dei contesti archivistici". In: *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti*. Venezia: Associazione per l'Informatica Umanistica e la Cultura Digitale, 2024, pp. 411–416. URL: https://hdl.handle.net/11585/996145.

[78] Giagnolini, Lucia, Schimmenti, Andrea, Bonora, Paolo, and Tomasi, Francesca. "Expliciting Contexts: Semantic Knowledge Extraction from Traditional Archival Descriptions". In: *Umanistica Digitale* 9.20 (2025), pp. 115–144. DOI: 10.6092/issn.2532-8816/21229. URL: https://umanisticadigitale.unibo.it/article/view/21229.

[79] González-Gallardo, Carlos-Emiliano, Boros, Emanuela, Giamphy, Edward, Hamdi, Ahmed, Moreno, José G., and Doucet, Antoine. "Injecting Temporal-Aware Knowledge in Historical Named Entity Recognition". In: *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*. Dublin, Ireland: Springer-Verlag, 2023, pp. 377–393. ISBN: 978-3-031-28243-0. DOI: 10.1007/978-3-031-28244-7_24. URL: https://doi.org/10.1007/978-3-031-28244-7_24.

[80] Google. *Google Scholar*. https://scholar.google.com/. Accessed: 2025-09-18.

[81] Graciotti, Arianna, Piano, Leonardo, Lazzari, Nicolas, Daga, Enrico, Tripodi, Rocco, Presutti, Valentina, and Pompianu, Livio. " KE-MHISTO: Towards a Multilingual Historical Knowledge Extraction Benchmark for Addressing the Long-Tail Problem ". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Che, Wanxiang, Nabende, Joyce, Shutova, Ekaterina, and Pilehvar, Mohammad Taher. Association for Computational Linguistics, 2025, pp. 20316–20339. ISBN: 979-8-89176-256-5. DOI: 10.18653/v1/2025.findings-acl.1042. URL: https://aclanthology.org/2025.findings-acl.1042/.

[82] Groth, P.T., Gibson, A, and Velterop, J. "The anatomy of a nanopublication". English. In: *Information Services and Use* 30 (2010), pp. 51–56. ISSN: 0167-5265. DOI: 10.3233/ISU-2010-0613.

[83] Gu, Jiawei, Jiang, Xuhui, Shi, Zhichao, Tan, Hexiang, Zhai, Xuehao, Xu, Chengjin, Li, Wei, Shen, Yinghan, Ma, Shengjie, Liu, Honghao, Wang, Saizhuo, Zhang, Kun, Wang, Yuanzhuo, Gao, Wen, Ni, Lionel, and Guo, Jian. *A Survey on LLM-as-a-Judge*. 2025. arXiv: 2411.15594 [cs.CL]. URL: https://arxiv.org/abs/2411.15594.

[84] Guarino, Nicola. "Formal ontology, conceptual analysis and knowledge representation". In: *International Journal of Human-Computer Studies* 43.5 (1995), pp. 625–640. ISSN: 1071-5819. DOI: https://doi.org/10.1006/ijhc.1995.1066. URL: https://www.sciencedirect.com/science/article/pii/S107158198571066X.

[85] Haider, S. *Verzeichnis Der Den Oberösterreichischen Raum Betreffenden Gefälschten, Manipulierten Oder Verdächtigten Mittelalterlichen Urkunden*. Tech. rep. Oberösterreichisches Landesarchiv, 2022.

[86] Hamborg, F., Donnay, K., and Gipp, B. "Towards Target-Dependent Sentiment Classification in News Articles". In: *Diversity, Divergence, Dialogue. iConference 2021.* Ed. by Toeppe, K., Yan, H., and Chu, S.K.W. Vol. 12646. Lecture Notes in Computer Science. Springer, 2021, pp. 157–169. DOI: 10.1007/978-3-030-71305-8_12.

[87] Han, Daniel, Han, Michael, and al., et. *Unsloth.* 2023. URL: http://github.com/unslothai/unsloth.

[88] Harpring, Patricia. "Development of the Getty Vocabularies: AAT, TGN, ULAN, and CONA". In: *Art Doc.* 29.1 (2010), pp. 67–72.

[89] Härtel, R. "Il Falso Documento Del Conte Giovanni Di Moggio (875)". In: *Mueç. Societât Filologjiche Furlane/Società Filologica Friulana, XCIV Congrès.* Ed. by Pugnetti, G. and Lucci, B. 2017, pp. 247–252.

[90] Hartig, Olaf. "Foundations of RDF*and SPARQL*:(An alternative approach to statement-level metadata in RDF)". In: *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web.* Ed. by Reutter, Juan L. and Srivastava, Divesh. Vol. 1912. CEUR Workshop Proceedings. Juan Reutter, Divesh Srivastava. Montevideo, Uruguay: CEUR-WS.org, 2017. URL: http://ceur-ws.org/Vol-1912/paper12.pdf.

[91] He, Jie, Yang, Yijun, Long, Wanqiu, Xiong, Deyi, Gutierrez Basulto, Victor, and Pan, Jeff Z. "Evaluating and Improving Graph to Text Generation with Large Language Models". In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).* Ed. by Chiruzzo, Luis, Ritter, Alan, and Wang, Lu. Association for Computational Linguistics, 2025, pp. 10219–10244. ISBN: 979-8-89176-189-6. DOI: 10.18653/v1/2025.naacl-long.513. URL: https://aclanthology.org/2025.naacl-long.513/.

[92] Higgins, Sarah. "The DCC Curation Lifecycle Model". English. In: *International Journal of Digital Curation* 3.1 (2008), pp. 134–140. ISSN: 1746-8256. DOI: 10.2218/ijdc.v3i1.48.

[93] Holtzman, Ari, Buys, Jan, Du, Li, Forbes, Maxwell, and Choi, Yejin. "The Curious Case of Neural Text Degeneration". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=rygGQyrFvH.

[94] Honnibal, Matthew, Montani, Ines, Van Landeghem, Sofie, and Boyd, Adriane. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. DOI: 10.5281/zenodo.1212303. URL: https://doi.org/10.5281/zenodo.1212303.

[95] Hotho, Andreas, Martinez-Rodriguez, Jose L., Hogan, Aidan, and Lopez-Arevalo, Ivan. "Information extraction meets the Semantic Web: A survey". In: *Semant. Web* 11.2 (2020). Place: NLD Publisher: IOS Press, pp. 255–335. ISSN: 1570-0844. DOI: 10.3233/SW-180333. URL: https://doi.org/10.3233/SW-180333.

[96] Huguet Cabot, Pere-Lluis. " From text to knowledge: multilingual information extraction for knowledge graph construction ". PhD thesis. PhD thesis. Università degli Studi di Roma "La Sapienza", 2025. URL: https://hdl.handle.net/11573/1732651.

[97] Hyvönen, Eero. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Synthesis Lectures on Data, Semantics, and Knowledge. Cham: Springer, 2012, pp. xiv+145. ISBN: 978-3-031-79438-4. DOI: 10.1007/978-3-031-79438-4.

[98] Hyvönen, Eero. "Digital humanities on the Semantic Web: Sampo model and portal series". In: *Semantic Web* 14.4 (2023), pp. 729–744. DOI: 10.3233/SW-223034. eprint: https://doi.org/10.3233/SW-223034. URL: https://doi.org/10.3233/SW-223034.

[99] Hyvönen, Eero, Leskinen, Petri, Tamper, Minna, Rantala, Heikki, Ikkala, Esko, Tuominen, Jouni, and Keravuori, Kirsi. "BiographySampo - Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research". In: *The Semantic Web. ESWC 2019*. Ed. by Hitzler, Pascal, Fernández, Miriam, Janowicz, Krzysztof, Zaveri, Amrapali, Gray, Alasdair J.G., Lopez, Vanessa, Haller, Armin, and Hammar, Karl. Springer-Verlag, 2019, pp. 574–589. DOI: 10.1007/978-3-030-21348-0_37. URL: https://doi.org/10.1007/978-3-030-21348-0_37.

[100] IEEE. *IEEE Xplore Digital Library*. https://ieeexplore.ieee.org/. Accessed: 2025-09-18.

[101] IOS Press. *Semantic Web Journal*. https://www.semantic-web-journal.net/. Accessed: 2025-09-18.

[102] Iqbal, Rizwan, Murad, Masrah Azrifah Azmi, Mustapha, A., and Sharef, Nurfadhlina Mohd. "An Analysis of Ontology Engineering Methodologies: A Literature Review". In: *Research Journal of Applied Sciences, Engineering and Technology* 6 (2013), pp. 2993–3000. URL: https://api.semanticscholar.org/CorpusID:16954591.

[103] *International Semantic Web Conference*. https://iswc2024.semanticweb.org/. Accessed: 2025-09-18.

[104] Jain, Nitisha, Sierra-Múnera, Alejandro, Ehmueller, Jan, and Krestel, Ralf. "Generation of training data for named entity recognition of artworks". In: *Semantic Web* 14.2 (2022), pp. 239–260. DOI: 10.3233/SW-223177. eprint: https://journals.sagepub.com/doi/pdf/10.3233/SW-223177. URL: https://journals.sagepub.com/doi/abs/10.3233/SW-223177.

[105] Järvelin, Kalervo and Kekäläinen, Jaana. "IR evaluation methods for retrieving highly relevant documents". In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '00. Athens, Greece: Association for Com-

puting Machinery, 2000, pp. 41–48. ISBN: 1581132263. DOI: 10.1145/345508.345545. URL: https://doi.org/10.1145/345508.345545.

[106] Johnson, Jeff, Douze, Matthijs, and Jégou, Hervé. "Billion-Scale Similarity Search with GPUs". In: *IEEE Transactions on Big Data* 7.3 (2021), pp. 535–547. DOI: 10.1109/TBDATA.2019.2921572.

[107] Journal, Semantic Web. *Special Issue on Large Language Models, Generative AI and Knowledge Graphs*. https://www.semantic-web-journal.net/blog/special-issue-large-language-models-generative-ai-and-knowledge-graphs. Accessed: 2025-01-21. 2024.

[108] Jurafsky, Dan and Martin, James H. *Speech and Language Processing*. 2nd ed. Prentice Hall, 2009. ISBN: 978-0131873216.

[109] Klie, Jan-Christoph, Bugert, Michael, Boullosa, Beto, Eckart de Castilho, Richard, and Gurevych, Iryna. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation". In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Ed. by Zhao, Dongyan. Association for Computational Linguistics, 2018, pp. 5–9. URL: https://aclanthology.org/C18-2002/.

[110] Koho, Mikko, Leskinen, Petri, and Hyvönen, Eero. " Integrating Historical Person Registers as Linked Open Data in the WarSampo Knowledge Graph ". In: *Semantic Systems. In the Era of Knowledge Graphs*. Ed. by Blomqvist, Eva, Groth, Paul, Boer, Victor de, Pellegrini, Tassilo, Alam, Mehwish, Käfer, Tobias, Kieseberg, Peter, Kirrane, Sabrina, Meroño-Peñuela, Albert, and Pandit, Harshvardhan J. Springer International Publishing, 2020, pp. 118–126. ISBN: 978-3-030-59833-4.

[111] Koolen, Marijn, Bogers, Toine, Gäde, Maria, Hall, Mark, Hendrickx, Iris, Huurdeman, Hugo, Kamps, Jaap, Skov, Mette, Verberne, Suzan, and Walsh, David. "Overview of the CLEF 2016 Social Book Search Lab". In: *Experimental IR Meets Multilinguality, Multimodality, and In-*

teraction. Ed. by Fuhr, Norbert, Quaresma, Paulo, Gonçalves, Teresa, Larsen, Birger, Balog, Krisztian, Macdonald, Craig, Cappellato, Linda, and Ferro, Nicola. Cham: Springer International Publishing, 2016, pp. 351–370. ISBN: 978-3-319-44564-9.

[112] Krippendorff, Klaus. *Content Analysis: An Introduction to Its Methodology*. Fourth. SAGE Publications, Inc., 2019. DOI: 10.4135/9781071878781. URL: https://methods.sagepub.com/book/mono/content-analysis-4e/toc.

[113] Kuculo, Tin, Abdollahi, Sara, and Gottschalk, Simon. " Transformer-Based Architectures versus Large Language Models in Semantic Event Extraction: Evaluating Strengths and Limitations ". In: *Semantic Web* (2024). Tracking #: 3673-4887, Major Revision. URL: https://github.com/t-kuculo/T-SEE.

[114] Kusnick, Jakob, Mayr, Eva, Seirafi, Kasra, Beck, Samuel, Liem, Johannes, and Windhager, Florian. " Every Thing Can Be a Hero! Narrative Visualization of Person, Object, and Other Biographies ". In: *Informatics* 11.2 (2024), p. 26. ISSN: 2227-9709. DOI: 10.3390/informatics11020026. URL: https://www.mdpi.com/2227-9709/11/2/26.

[115] Leskinen, Petri and Hyvönen, Eero. "Extracting Genealogical Networks of Linked Data from Biographical Texts". In: *The Semantic Web: ESWC 2019 Satellite Events: ESWC 2019 Satellite Events, Portorož, Slovenia, June 2–6, 2019, Revised Selected Papers*. Portoroz, Slovenia: Springer-Verlag, 2019, pp. 121–125. ISBN: 978-3-030-32326-4. DOI: 10.1007/978-3-030-32327-1_24.

[116] Lewis, Patrick, Perez, Ethan, Piktus, Aleksandra, Petroni, Fabio, Karpukhin, Vladimir, Goyal, Naman, Küttler, Heinrich, Lewis, Mike, Yih, Wen-tau, Rocktäschel, Tim, Riedel, Sebastian, and Kiela, Douwe. "Retrieval-augmented generation for knowledge-intensive NLP tasks".

In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[117] Li, Bangzheng, Yin, Wenpeng, and Chen, Muhao. "Ultra-fine Entity Typing with Indirect Supervision from Natural Language Inference". In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 607–622. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00479.

[118] Li, Bohan, Hou, Yutai, and Che, Wanxiang. "Data augmentation approaches in natural language processing: A survey". In: *AI Open* 3 (2022), pp. 71–90. ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2022.03.001. URL: https://www.sciencedirect.com/science/article/pii/S2666651022000080.

[119] Li, Ziyan, Wang, Haofen, and Zhang, Wenqiang. "Translational relation embeddings for multi-hop knowledge base question answering". In: *Journal of Web Semantics* 74 (2022), p. 100723. ISSN: 1570-8268. DOI: https://doi.org/10.1016/j.websem.2022.100723. URL: https://www.sciencedirect.com/science/article/pii/S1570826822000154.

[120] Lisena, Pasquale, Ehrhart, Thibault, and Troncy, Raphaël. "How to Embed Large but Incomplete Knowledge Graphs in the Culture Heritage Sector: Lessons Learned from Odeuropa". In: *Handbook on Neurosymbolic AI and Knowledge Graphs* (2025).

[121] Lisena, Pasquale, Erp, Marieke van, Bembibre, Cecilia, and Leemans, Inger. " Data mining and knowledge graphs as a backbone for advanced olfactory experiences ". In: *STT 2021: Smell, Taste, and Temperature Interfaces workshop at CHI 2021, 7–9 May 2021, Yokohama, Japan (Virtual Conference)
* ed. by ACM. 2021.

[122] Lisena, Pasquale, Schwabe, Daniel, Erp, Marieke van, Troncy, Raphaël, Tullett, William, Leemans, Inger, Marx, Lizzie, and Ehrich, Sofia Colette. “ Capturing the Semantics of Smell: The Odeuropa Data Model for Olfactory Heritage Information ”. In: *The Semantic Web*. Ed. by Groth, Paul, Vidal, Maria-Esther, Suchanek, Fabian, Szekley, Pedro, Kapanipathi, Pavan, Pesquita, Catia, Skaf-Molli, Hala, and Tamper, Minna. Springer International Publishing, 2022, pp. 387–405. ISBN: 978-3-031-06981-9.

[123] Liu, Yang, Iter, Dan, Xu, Yichong, Wang, Shuohang, Xu, Ruochen, and Zhu, Chenguang. “G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Bouamor, Houda, Pino, Juan, and Bali, Kalika. Association for Computational Linguistics, 2023, pp. 2511–2522. DOI: 10.18653/v1/2023.emnlp-main.153. URL: https://aclanthology.org/2023.emnlp-main.153/.

[124] Liu, Zhaoming. “Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies”. In: *Journal of Transcultural Communication* 3.2 (2023), pp. 224–244. DOI: doi:10.1515/jtc-2023-0019. URL: https://doi.org/10.1515/jtc-2023-0019.

[125] *Knowledge-Base Construction from Pretrained Language Models Challenge*. https://lm-kbc.github.io/challenge2025/. Accessed: 2025-09-18.

[126] LMKBC. *Knowledge-Base Construction from Pretrained Language Models Challenge*. https://lm-kbc.github.io/challenge2025/. Accessed: 2025-01-21. 2025.

[127] Long, Lin, Wang, Rui, Xiao, Ruixuan, Zhao, Junbo, Ding, Xiao, Chen, Gang, and Wang, Haobo. “On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Ku, Lun-Wei, Mar-

tins, Andre, and Srikumar, Vivek. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 11065–11082. DOI: 10.18653/v1/2024.findings-acl.658. URL: https://aclanthology.org/2024.findings-acl.658/.

[128] Luo, Haoran, E, Haihong, Yang, Yuhao, Yao, Tianyu, Guo, Yikai, Tang, Zichen, Zhang, Wentai, Peng, Shiyao, Wan, Kaiyang, Song, Meina, Lin, Wei, Zhu, Yifan, and Tuan, Luu Anh. " Text2NKG: Fine-Grained N-ary Relation Extraction for N-ary relational Knowledge Graph Construction ". In: *Advances in Neural Information Processing Systems*. Ed. by Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. Vol. 37. Curran Associates, Inc., 2024, pp. 27417–27439. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/305b2288122d46bf0641bdd86c9a7921-Paper-Conference.pdf.

[129] Luthra, Mrinalini and Eskevich, Maria. "Data-Envelopes for Cultural Heritage: Going beyond Datasheets". In: *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies LREC-COLING 2024*. Ed. by Siegert, Ingo and Choukri, Khalid. Torino, Italia: ELRA and ICCL, 2024, pp. 52–65. URL: https://aclanthology.org/2024.legal-1.9/.

[130] Luthra, Mrinalini, Todorov, Konstantin, Jeurgens, Charles, and Colavizza, Giovanni. "Unsilencing colonial archives via automated entity recognition". In: *Journal of Documentation* 80.5 (2023), pp. 1080–1105. ISSN: 0022-0418. DOI: 10.1108/JD-02-2022-0038. eprint: https://www.emerald.com/jd/article-pdf/80/5/1080/9601322/jd-02-2022-0038.pdf. URL: https://doi.org/10.1108/JD-02-2022-0038.

[131] Mausam, Mausam. "Open information extraction systems and downstream applications". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI'16. New York, New York, USA: AAAI Press, 2016, pp. 4074–4077. ISBN: 9781577357704.

[132] Maynard, Diana, Bontcheva, Kalina, and Augenstein, Isabelle. *Natural Language processing for the Semantic Web*. Synthesis Lectures on the semantic web: theory and technology 15. Morgan & Claypool Publishers, 2017. DOI: https://doi.org/10.1007/978-3-031-79474-2.

[133] Menini, Stefano. "Semantic Frame Extraction in Multilingual Olfactory Events". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Calzolari, Nicoletta, Kan, Min-Yen, Hoste, Veronique, Lenci, Alessandro, Sakti, Sakriani, and Xue, Nianwen. Torino, Italia: ELRA and ICCL, 2024, pp. 14622–14627. URL: https://aclanthology.org/2024.lrec-main.1273.

[134] Menis Mastromichalakis, Orfeas, Liartis, Jason, Rose, Kristina, Isaac, Antoine, and Stamou, Giorgos. "Don't Erase, Inform! Detecting and Contextualizing Harmful Language in Cultural Heritage Collections". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Che, Wanxiang, Nabende, Joyce, Shutova, Ekaterina, and Pilehvar, Mohammad Taher. Association for Computational Linguistics, 2025, pp. 21836–21850. ISBN: 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.1060. URL: https://aclanthology.org/2025.acl-long.1060/.

[135] Meyer, Lars-Peter, Stadler, Claus, Frey, Johannes, Radtke, Norman, Junghanns, Kurt, Meissner, Roy, Dziwis, Gordian, Bulert, Kirill, and Martin, Michael. "LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT". In: *First Working Conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow*. Ed. by Zinke-Wehlmann, Christian and Friedrich, Julia. Wiesbaden: Springer Fachmedien Wiesbaden, 2024, pp. 103–115. ISBN: 978-3-658-43705-3.

[136] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?" In: *Proceedings of the 2022 Con-*

ference on Empirical Methods in Natural Language Processing. Ed. by Goldberg, Yoav, Kozareva, Zornitsa, and Zhang, Yue. Association for Computational Linguistics, 2022, pp. 11048–11064. URL: https://aclanthology.org/2022.emnlp-main.759.

[137] Morales Tirado, Alba, Carvalho, Jason, Mulholland, Paul, and Daga, Enrico. " Musical Meetups: a Knowledge Graph approach for Historical Social Network Analysis ". In: *ESWC 2023 Workshops and Tutorials. Semantic Methods for Events and Stories (SEMMES)*. Ed. by Alam, Mehwish, Trojahn, Cassia, Hertling, Sven, Pesquita, Catia, Aebeloe, Christian, Aras, Hidir, Azzam, Amr, Cano, Juan, John Domingue, Gottschalk, Simon, Hartig, Olaf, Hose, Katja, Sabrina Kirrane, Lisena, Pasquale, Osborne, Francesco, Rohde, Philipp, Luc Steels, Taelman, Ruben, Third, Aisling, Tiddi, Ilaria, and Rima Türker. Vol. 3443. ISSN: 1613-0073. CEUR Workshop Proceedings (CEUR-WS.org), 2023. URL: https://ceur-ws.org/Vol-3443/ESWC%5C%5f2023%5C%5fSEMMES%5C%5fMeetups-CR.pdf.

[138] Nuzzolese, Andrea Giovanni, Gentile, Anna Lisa, Presutti, Valentina, Gangemi, Aldo, Garigliotti, Darío, and Navigli, Roberto. "Open knowledge extraction challenge". In: *Semantic Web Evaluation Challenges: Second SemWebEval Challenge at ESWC 2015, Portorož, Slovenia, May 31-June 4, 2015, Revised Selected Papers*. Springer. 2015, pp. 3–15.

[139] Olderr, Steven. *Symbolism: A Comprehensive Dictionary*. 2nd ed. London: McFarland, 2012.

[140] Omura, Kazumasa, Cheng, Fei, and Kurohashi, Sadao. "An Empirical Study of Synthetic Data Generation for Implicit Discourse Relation Recognition". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Calzolari, Nicoletta, Kan, Min-Yen, Hoste, Veronique, Lenci, Alessandro, Sakti, Sakriani, and Xue,

Nianwen. Torino, Italia: ELRA and ICCL, 2024, pp. 1073–1085. URL: https://aclanthology.org/2024.lrec-main.96/.

[141] Ondraszek, Sarah Rebecca, Petri, Grischka, Blumenthal, Ulrike, Dieckmann, Lisa, Posthumus, Etienne, and Sack, Harald. " eXtreme Design for Ontological Engineering in the Digital Humanities with Viewsari, a Knowledge Graph of Giorgio Vasari's The Lives ". In: *SemDHESWC*. 2024. URL: https://api.semanticscholar.org/CorpusID:271408170.

[142] Ondraszek, Sarah Rebecca, Sack, Harald, and Posthumus, Etienne. " Viewsari: New Perspectives on Historical Network Analysis in Giorgio Vasari's The Lives Using Knowledge Graphs ". In: *Historical Network Research, Zenodo, Lausanne* (2024). URL: https://fizweb-p.fiz-karlsruhe.de/sites/default/files/FIZ/Dokumente/Forschung/ISE/Publications/Conferences-Workshops/Ondraszek-HNR-2024.pdf (visited on 09/20/2024).

[143] Ontotext. *GraphDB: Semantic Database*. Accessed: March 2025. 2024. URL: https://www.ontotext.com/products/graphdb/.

[144] *OpenCitations*. https://search.opencitations.net/. Accessed: 2025-09-18.

[145] Orlando, Riccardo, Huguet Cabot, Pere-Lluís, Barba, Edoardo, and Navigli, Roberto. "Retrieve, Read and LinK: Fast and Accurate Entity Linking and Relation Extraction on an Academic Budget". In: *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024.

[146] Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing. "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[147] Pasqual, Valentina. "The Critical Inquiry in Humanities Knowledge Graphs: Challenges, Methods, Innovations". PhD thesis. alma, 2025.

[148] Pasqual, Valentina, Pedretti, Carlo Teo, Schimmenti, Andrea, Tomasi, Francesca, and Vitali, Fabio. "Non-representational approaches to visualise complex information in the Cultural Heritage domain". In: *Digital Humanities 2023: Book of Abstracts*. Ed. by Scholger, Walter, Vogeler, Georg, Tasovac, Toma, Baillot, Anne, and Helling, Patrick. 2023, pp. 376–378. DOI: 10.5281/zenodo.7961821. URL: https://hdl.handle.net/11585/945552.

[149] Peeters, Leo. "Historiciteit en chronologie in Van den vos Reynaerde". In: *Spektator. Tijdschrift voor Neerlandistiek* 3 (1973), pp. 157–175. URL: https://www.dbnl.org/tekst/_spe011197301_01/_spe011197301_01_0022.php.

[150] Pellegrino, Maria Angela, Scarano, Vittorio, and Spagnuolo, Carmine. "Move cultural heritage knowledge graphs in everyone's pocket". In: *Semantic Web* 14.2 (2022), pp. 323–359. DOI: 10.3233/SW-223117. eprint: https://journals.sagepub.com/doi/pdf/10.3233/SW-223117. URL: https://journals.sagepub.com/doi/abs/10.3233/SW-223117.

[151] Peroni, Silvio. "A Simplified Agile Methodology for Ontology Development". In: *OWL: Experiences and Directions – Reasoner Evaluation: 13th International Workshop, OWLED 2016, and 5th International Workshop, ORE 2016, Bologna, Italy, November 20, 2016, Revised Selected Papers*. Bologna, Italy: Springer-Verlag, 2016, pp. 55–69. ISBN: 978-3-319-54626-1. DOI: 10.1007/978-3-319-54627-8_5. URL: https://doi.org/10.1007/978-3-319-54627-8_5.

[152] Petroni, Fabio, Rocktäschel, Tim, Riedel, Sebastian, Lewis, Patrick, Bakhtin, Anton, Wu, Yuxiang, and Miller, Alexander. "Language Models as Knowledge Bases?" In: *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Ed. by Inui, Kentaro, Jiang, Jing, Ng, Vincent, and Wan, Xiaojun. Association for Computational Linguistics, 2019, pp. 2463–2473. DOI: 10.18653/v1/D19-1250. URL: https://aclanthology.org/D19-1250/.

[153] Piotrowski, M. "Uncertainty as Unavoidable Good". In: *Universität Bielefeld, Center for Uncertainty Studies (CeUS)* 5 (2023), p. 10. DOI: 10.4119/unibi/2983506. URL: https://pub.uni-bielefeld.de/record/2983506.

[154] Piotrowski, M. and Neuwirth, M. "Prospects for computational hermeneutics". In: *Proceedings of the 9th AIUCD Annual Conference*. 2020. URL: http://amsacta.unibo.it/6316/.

[155] Piotrowski, Michael. "Accepting and Modeling Uncertainty". In: *Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graph-basierten Modellierung von Unsicherheiten*. Ed. by Kuczera, Andreas, Wübbena, Thorsten, and Kollatz, Thomas. Sonderbände 4. Wolfenbüttel: Zeitschrift für digitale Geisteswissenschaften, 2019. DOI: 10.17175/sb004_006a. URL: https://zfdg.de/sb004_006 (visited on 01/15/2025).

[156] Pontiki, Maria, Galanis, Dimitris, Pavlopoulos, John, Papageorgiou, Harris, Androutsopoulos, Ion, and Manandhar, Suresh. "SemEval-2014 Task 4: Aspect Based Sentiment Analysis". In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Ed. by Nakov, Preslav and Zesch, Torsten. Dublin, Ireland: Association for Computational Linguistics, 2014, pp. 27–35. DOI: 10.3115/v1/S14-2004. URL: https://aclanthology.org/S14-2004/.

[157] Popović, Maja. "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine*

Translation. Association for Computational Linguistics, 2015, pp. 392–395.

[158] Presutti, Valentina, Daga, Enrico, Gangemi, Aldo, and Blomqvist, Eva. "eXtreme design with content ontology design patterns". In: *Proceedings of the 2009 International Conference on Ontology Patterns - Volume 516*. WOP'09. Washington DC: CEUR-WS.org, 2009, pp. 83–97.

[159] Qin, Yujia, Hu, Shengding, Lin, Yankai, Chen, Weize, Ding, Ning, Cui, Ganqu, Zeng, Zheni, Zhou, Xuanhe, Huang, Yufei, Xiao, Chaojun, Han, Chi, Fung, Yi Ren, Su, Yusheng, Wang, Huadong, Qian, Cheng, Tian, Runchu, Zhu, Kunlun, Liang, Shihao, Shen, Xingyu, Xu, Bokai, Zhang, Zhen, Ye, Yining, Li, Bowen, Tang, Ziwei, Yi, Jing, Zhu, Yuzhang, Dai, Zhenning, Yan, Lan, Cong, Xin, Lu, Yaxi, Zhao, Weilin, Huang, Yuxiang, Yan, Junxi, Han, Xu, Sun, Xian, Li, Dahai, Phang, Jason, Yang, Cheng, Wu, Tongshuang, Ji, Heng, Li, Guoliang, Liu, Zhiyuan, and Sun, Maosong. "Tool Learning with Foundation Models". In: *ACM Comput. Surv.* 57.4 (2024). ISSN: 0360-0300. DOI: 10.1145/3704435. URL: https://doi.org/10.1145/3704435.

[160] Radchenko, Vladyslav and Drushchak, Nazarii. "Improving Named Entity Recognition for Low-Resource Languages Using Large Language Models: A Ukrainian Case Study". In: *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*. Ed. by Romanyshyn, Mariana. Association for Computational Linguistics, 2025, pp. 27–35. ISBN: 979-8-89176-269-5. DOI: 10.18653/v1/2025.unlp-1.3. URL: https://aclanthology.org/2025.unlp-1.3/.

[161] Ranjgar, Babak, Sadeghi-Niaraki, Abolghasem, Shakeri, Maryam, Rahimi, Fatema, and Choi, Soo-Mi. " Cultural Heritage Information Retrieval: Past, Present, and Future Trends ". In: *IEEE Access* 12 (2024), pp. 42992–43026. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3374769. URL: https://ieeexplore.ieee.org/document/10463018 (visited on 09/17/2024).

[162] Reforgiato Recupero, Diego, Presutti, Valentina, Consoli, Sergio, Gangemi, Aldo, and Nuzzolese, Andrea Giovanni. "Sentilo: Frame-Based Sentiment Analysis". In: *Cognitive Computation* 7.2 (2015), pp. 211–225. ISSN: 1866-9964. DOI: 10.1007/s12559-014-9302-z. URL: https://doi.org/10.1007/s12559-014-9302-z.

[163] Regino, Andre, Rossanez, Anderson, Torres, Ricardo da Silva, and Reis, Julio Cesar dos. " A Systematic Literature Review on RDF Triple Generation from Natural Language Texts ". In: *Semantic Web* (2025). Accepted for publication, Tracking #3894-5108.

[164] Ringwald, Célian, Gandon, Fabien, Faron, Catherine, Michel, Franck, and Akl, Hanna Abi. "12 Shades of RDF: Impact of Syntaxes on Data Extraction with Language Models". In: *The Semantic Web: ESWC 2024 Satellite Events: Hersonissos, Crete, Greece, May 26–30, 2024, Proceedings, Part I*. Hersonissos, Greece: Springer-Verlag, 2025, pp. 81–91. ISBN: 978-3-031-78951-9. DOI: 10.1007/978-3-031-78952-6_8. URL: https://doi.org/10.1007/978-3-031-78952-6_8.

[165] Ristoski, Petar, Lin, Zhizhong, and Zhou, Qunzhi. "KG-ZESHEL: Knowledge Graph-Enhanced Zero-Shot Entity Linking". In: *Proceedings of the 11th Knowledge Capture Conference*. K-CAP '21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 49–56. ISBN: 9781450384575. DOI: 10.1145/3460210.3493549. URL: https://doi.org/10.1145/3460210.3493549.

[166] Rosales-Méndez, Henry, Hogan, Aidan, and Poblete, Barbara. "Fine-Grained Evaluation for Entity Linking". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Inui, Kentaro, Jiang, Jing, Ng, Vincent, and Wan, Xiaojun. Hong Kong, China: Association for Computational

Linguistics, 2019, pp. 718–727. DOI: 10.18653/v1/D19-1066. URL: https://aclanthology.org/D19-1066/.

[167] Russell, Stuart J. and Norvig, Peter. *Artificial Intelligence: A Modern Approach*. 2nd ed. Pearson Education, 2003. ISBN: 0137903952.

[168] Saeidi, Marzieh, Bouchard, Guillaume, Liakata, Maria, and Riedel, Sebastian. "SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Ed. by Matsumoto, Yuji and Prasad, Rashmi. Osaka, Japan: The COLING 2016 Organizing Committee, 2016, pp. 1546–1556. URL: https://aclanthology.org/C16-1146/.

[169] Sahbi, Aya, Alec, Céline, and Beust, Pierre. "Automatic Ontology Population from Textual Advertisements: LLM vs. Semantic Approach". In: *Procedia Computer Science* 246 (2024). 28th International Conference on Knowledge Based and Intelligent Information & Engineering Systems (KES 2024), pp. 3083–3092. DOI: 10.1016/j.procs.2024.09.364. URL: https://doi.org/10.1016/j.procs.2024.09.364.

[170] Santini, Cristian. " Combining language models for knowledge extraction from Italian TEI editions ". In: *Frontiers in Computer Science* 6 (2024). Section: Human-Media Interaction, p. 1472512. DOI: 10.3389/fcomp.2024.1472512. URL: https://doi.org/10.3389/fcomp.2024.1472512.

[171] Santini, Cristian, Melosi, Laura, and Frontoni, Emanuele. *Named Entity Recognition in Historical Italian: The Case of Giacomo Leopardi's Zibaldone*. 2025. arXiv: 2505.20113 [cs.CL]. URL: https://arxiv.org/abs/2505.20113.

[172] Santini, Cristian, Tan, Mary Ann, Tietz, Tabea, Bruns, Oleksandra, Posthumus, Etienne, and Sack, Harald. " 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) Berlin, Germany, Sept. 19th-23rd, 2022. Ed.: A. Paschke*. 3rd Conference on Digital Curation Technologies. Qurator 2022 (Berlin, Deutschland, Sept. 19–23, 2022). Vol. 3234. CEUR Workshop Proceedings. CEUR-WS.org, 2022.

[173] Sartini, B., Baroncini, S., Erp, M. van, Tomasi, F., and Gangemi, A. "ICON: An Ontology for Comprehensive Artistic Interpretations". In: *J. Comput. Cult. Herit.* 16.3 (2023), pp. 59–76. DOI: 10.1145/3594724.

[174] Sartini, Bruno, Erp, Marieke van, and Gangemi, Aldo. " Marriage is a Peach and a Chalice: Modelling Cultural Symbolism on the Semantic Web ". In: *Proceedings of the 11th Knowledge Capture Conference*. K-CAP '21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 201–208. ISBN: 9781450384575. DOI: 10.1145/3460210.3493552. URL: https://doi.org/10.1145/3460210.3493552.

[175] Schick, Timo, Dwivedi-Yu, Jane, Dessi, Roberto, Raileanu, Roberta, Lomeli, Maria, Hambro, Eric, Zettlemoyer, Luke, Cancedda, Nicola, and Scialom, Thomas. "Toolformer: Language Models Can Teach Themselves to Use Tools". In: *Advances in Neural Information Processing Systems*. Ed. by Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. Vol. 36. Curran Associates, Inc., 2023, pp. 68539–68551. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf.

[176] Schimmenti, Andrea. *Book Review ABSA+ET Dataset*. 2025. DOI: 10.57967/hf/4915. URL: https://huggingface.co/datasets/aschimmenti2/absa_llama.

[177] Schimmenti, Andrea. *Llama3.1 8B Fine Tuned model for ABSA+ET*. 2025. DOI: 10.57967/hf/4916. URL: https://huggingface.co/aschimmenti2/llama-absa-model.

[178] Schimmenti, Andrea, De Giorgis, Stefano, Vitali, Fabio, and Erp, Marieke van. "Old Reviews, New Aspects: Aspect Based Sentiment Analysis and Entity Typing for Book Reviews with LLMs". In: *Proceedings of the 5th Conference on Language, Data and Knowledge*. Ed. by Alam, Mehwish, Tchechmedjiev, Andon, Gracia, Jorge, Gromann, Dagmar, Buono, Maria Pia di, Monti, Johanna, and Ionov, Maxim. Unior Press, 2025, pp. 266–276. ISBN: 978-88-6719-333-2. URL: https://aclanthology.org/2025.ldk-1.27/.

[179] Schimmenti, Andrea, Pasqual, Valentina, Tomasi, Francesca, Vitali, Fabio, and Erp, Marieke van. "Structuring Authenticity Assessments on Historical Documents using LLMs". In: *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024*. 2024, pp. 463–468. DOI: 10.6092/unibo/amsacta/7927. URL: https://hdl.handle.net/11585/994558.

[180] Schimmenti, Andrea, Pasqual, Valentina, Vitali, Fabio, and Erp, Marieke van. "Knowledge Graphs Generation from Cultural Heritage Texts: Combining LLMs and Ontological Engineering for Scholarly Debates". In: *Journal of Documentation* (2025). Submitted for publication.

[181] Schlögl, Matthias, Tuominen, Jouni, Kesäniemi, Joonas, Leskinen, Petri, Sugimoto, Go, and Boer, Victor de. " The InTaVia Knowledge Graph – European National Biographical and Cultural Heritage Object Data ". In: *Semantic Web* (2025). Under review, Tracking #: 3851-5065. DOI: 10.5281/zenodo.5534542. URL: https://doi.org/10.5281/zenodo.5534542.

[182] *Workshop on Semantic Deep Learning*. https://semdh.github.io/. Accessed: 2025-09-18.

[183] Shyama, R., Wilson, I., Ginige, Athula, and Goonetillake, Jeevani. "Pattern-Based Prompting for Accurate Extraction of Ontology As-

sertional (A-box) Axioms Using LLMs for Ontology Population". In: *Procedia Computer Science* 270 (2025). 29th International Conference on Knowledge-Based and Intelligent Information Engineering Systems (KES 2025), pp. 1438–1447. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2025.09.265. URL: https://www.sciencedirect.com/science/article/pii/S1877050925029382.

[184] Springer. *SpringerLink*. https://link.springer.com/. Accessed: 2025-09-18.

[185] Stachowiak, Herbert. *Allgemeine Modelltheorie*. Springer Verlag, 1973. URL: https://archive.org/details/Stachowiak1973AllgemeineModelltheorie.

[186] Stramiglio, Alessandra, Schimmenti, Andrea, Pasqual, Valentina, Erp, Marieke van, Sovrano, Francesco, and Vitali, Fabio. *Explicit vs. Implicit Biographies: Evaluating and Adapting LLM Information Extraction on Wikidata-Derived Texts*. 2025. arXiv: 2509.14943 [cs.CL]. URL: https://arxiv.org/abs/2509.14943.

[187] Sugimoto, G., Daza, A., and Boer, V. de. " Closer Reading of RDF Generated by NLP on Wikipedia Biography: Comparative Analysis ". In: *Metadata and Semantic Research*. Communications in Computer and Information Science 2048 (2024). Ed. by Garoufallou, E. and Sartori, F. MTSR: Research Conference on Metadata and Semantics Research; Vol. 2023, pp. 41–54. DOI: 10.1007/978-3-031-65990-4_4.

[188] Tamašauskaitė, Gytė and Groth, Paul. " Defining a Knowledge Graph Development Process Through a Systematic Review ". In: *ACM Trans. Softw. Eng. Methodol.* 32.1 (2023), 27:1–27:40. ISSN: 1049-331X. DOI: 10.1145/3522586. URL: https://dl.acm.org/doi/10.1145/3522586 (visited on 09/20/2024).

[189] Tamper, Minna, Leal, Rafael, Sinikallio, Laura, Leskinen, Petri, Tuominen, Jouni, and Hyvönen, Eero. " Extracting Knowledge from Parliamentary Debates for Studying Political Culture and Language ".

In: *Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge (TEXT2KG 2022 and MK2022)*. Ed. by Tiwari, Sanju, Mihindukulasooriya, Nandana, Osborne, Francesco, et al. CEUR Workshop Proceedings. Peer reviewed. 2022, p. 10. URL: http://hdl.handle.net/10138/348159.

[190] Tamper, Minna, Leskinen, Petri, Ikkala, Esko, Oksanen, Arttu, Mäkelä, Eetu, Heino, Erkki, Tuominen, Jouni, Koho, Mikko, and Hyvönen, Eero. "AATOS – A Configurable Tool for Automatic Annotation". In: *Lecture Notes in Computer Science*. 2017, pp. 276–289. ISBN: 978-3-319-59887-1. DOI: 10.1007/978-3-319-59888-8_24.

[191] Tang, Alison D., Soulette, Cameron M., Baren, Marijke J. van, Hart, Kevyn, Hrabeta-Robinson, Eva, Wu, Catherine J., and Brooks, Angela N. "Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns". In: *Nature Communications* 11.1 (2020), p. 1438. DOI: 10.1038/s41467-020-15171-6. URL: https://doi.org/10.1038/s41467-020-15171-6.

[192] Tao, Yan, Viberg, Olga, Baker, Ryan S, and Kizilcec, René F. "Cultural bias and cultural alignment of large language models". In: *PNAS Nexus* 3.9 (2024), pgae346. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgae346. eprint: https://academic.oup.com/pnasnexus/article-pdf/3/9/pgae346/59151559/pgae346.pdf. URL: https://doi.org/10.1093/pnasnexus/pgae346.

[193] Terras, Melissa. "Digital humanities and digitised cultural heritage". English. In: *The Bloomsbury Handbook to the Digital Humanities*. Ed. by O'Sullivan, James. 1st ed. Bloomsbury Handbooks. Bloomsbury, 2022, pp. 255–266. ISBN: 9781350232112. DOI: 10.5040/9781350232143.ch-24.

[194] Testoni, Laura. "Digital curation e content curation: due risposte alla complessità dell'infosfera digitale che ci circonda, due sfide per i bibliotecari". In: *Bibliotime, Rivista elettronica per le biblioteche* XVI.1 (2013). ISSN 1128-3564.

[195] *International Workshop on Knowledge Graph Construction*. `https://text2kg.github.io/`. Accessed: 2025-09-18.

[196] Text2KG. *International Workshop on Knowledge Graph Construction (Text2KG)*. `https://aiisc.cse.sc.edu/text2kg2025/index.html`. Accessed: 2025-01-21. 2025.

[197] Thanapalasingam, Thiviyan, Krieken, Emile van, Bloem, Peter, and Groth, Paul. *IntelliGraphs: Datasets for Benchmarking Knowledge Graph Generation*. Issue: arXiv:2307.06698 arXiv: 2307.06698 [cs]. 2023. DOI: `10.48550/arXiv.2307.06698`. URL: `http://arxiv.org/abs/2307.06698` (visited on 09/20/2024).

[198] Tjong Kim Sang, Erik F. and De Meulder, Fien. "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://aclanthology.org/W03-0419/`.

[199] Tomasi, Francesca. "La preservazione del contenuto degli oggetti culturali: formalizzare la provenance". In: *BIBLIOTHECAE.IT* (2017).

[200] Tomasi, Francesca. *Vespasiano da Bisticci, Lettere. Knowledge Site 3.0.* Italian. Open Access, Creative Commons Attribution (CC BY). DH.arc, 2020, p. 200. ISBN: 9788898010110. DOI: `10.6092/unibo/amsacta/6852`. URL: `http://projects.dharc.unibo.it/vespasiano/`.

[201] Tomasi, Francesca. *Organizzare la conoscenza: Digital Humanities e Web semantico. Un percorso tra archivi, biblioteche e musei*. Biblioteconomia e Scienza dell'Informazione. Editrice Bibliografica, 2022,

p. 184. ISBN: 978-88-9357-357-3. DOI: 10.53134/9788893573573. URL: https://hdl.handle.net/11585/848605.

[202] Tomasi, Francesca. "Archival Finding Aids in Linked Open Data between Description and Interpretation". In: *JLIS.It* 14.3 (2023), pp. 134–146. DOI: https://doi.org/10.36253/jlis.it-557.

[203] Tonelli, Sara and Menini, Stefano. "FrameNet-like Annotation of Olfactory Information in Texts". In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by Degaetano-Ortlieb, Stefania, Kazantseva, Anna, Reiter, Nils, and Szpakowicz, Stan. Association for Computational Linguistics, 2021, pp. 11–20. DOI: 10.18653/v1/2021.latechclfl-1.2. URL: https://aclanthology.org/2021.latechclfl-1.2/.

[204] Valla, L. *The treatise of Lorenzo Valla on the Donation of Constantine*. Translated by Christopher Bush Coleman. Yale University Press, 2023. URL: https://www.gutenberg.org/ebooks/70092.

[205] Van Daele, Rik. "De robotfoto van de Reynaertdichter. Bricoleren met de overgeleverde wrakstukken: 'cisterciënzers', 'grafelijk hof' en 'Reynaertmaterie'". Dutch. In: *Tiecelijn: Nieuwsbrief voor Reynaerdofielen* 18.3 (2005), pp. 179–205. ISSN: 0775-9770. URL: https://repository.uantwerpen.be/docstore/d:irua:20388.

[206] Varagnolo, Davide, Melo, Dora, Rodrigues, Irene Pimenta, Rodrigues, Rui, and Couto, Paula. "Archives Metadata Text Information Extraction into CIDOC-CRM". In: *Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Ed. by Coenen, Frans, Fred, Ana, Aveiro, David, Dietz, Jan, Bernardino, Jorge, Masciari, Elio, and Filipe, Joaquim. Cham: Springer Nature Switzerland, 2023, pp. 195–216. ISBN: 978-3-031-43471-6.

[207] Vitali, Fabio and Pasqual, Valentina. "The Representation of Historical Uncertainties as the Outcome of Competing and Incompatible Certainties". In: *Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives*. Studi di archivistica, bibliografia, paleografia. Venice: Edizioni Ca' Foscari, 2025. ISBN: 978-88-6969-920-7. DOI: 10.30687/978-88-6969-919-1/003. URL: http://edizionicafoscari.it/it/edizioni4/libri/978-88-6969-919-1/the-representation-of-historical-uncertainties-as/.

[208] Wackers, Paul, Block, Elaine C., Dufournet, Jean, Goossens, Jan, Mann, Jill, Pastré, Jean-Marc, Schouwink, Wilfried, Stephenson, Roger, Subrenat, Jean, Suomela-Härmä, Elina, Daele, Rik van, and Varty, Kenneth. "Medieval French and Dutch Renardian Epics: Between Literature and Society". In: *Reynard the Fox: Cultural Metamorphoses and Social Engagement in the Beast Epic from the Middle Ages to the Present*. 1st ed. Berghahn Books, 2000, pp. 55–72. URL: http://www.jstor.org/stable/j.ctt1c0gkr9.9 (visited on 10/26/2025).

[209] Wackers, Paul W. M. "Wat staat er eigenlijk? Over het editeren van Van den vos Reynaerde". Dutch. In: *Tiecelijn* 29 (2016), pp. 98–118. URL: https://opac.regesta-imperii.de/id/2845838.

[210] Wadhwa, Somin, Amir, Silvio, and Wallace, Byron C. "Revisiting Relation Extraction in the era of Large Language Models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023, pp. 15566–15589. DOI: 10.18653/v1/2023.acl-long.868. URL: https://aclanthology.org/2023.acl-long.868.

[211] Wang, Shuhe, Sun, Xiaofei, Li, Xiaoya, Ouyang, Rongbin, Wu, Fei, Zhang, Tianwei, Li, Jiwei, Wang, Guoyin, and Guo, Chen. "GPT-NER: Named Entity Recognition via Large Language Models". In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Ed.

by Chiruzzo, Luis, Ritter, Alan, and Wang, Lu. Association for Computational Linguistics, 2025, pp. 4257–4275. ISBN: 979-8-89176-195-7. DOI: 10.18653/v1/2025.findings-naacl.239. URL: https://aclanthology.org/2025.findings-naacl.239/.

[212] Wang, Z., Xie, Q., Ding, Z., Feng, Y., and Xia, R. *Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study*. 2023. URL: https://api.semanticscholar.org/CorpusID:258048703.

[213] Web Sematics, Journal of. *Special Issue on Semantic Web and Information Retrieval*. https://www.sciencedirect.com/special-issue/10DWQV75Z4T. Accessed: 2025-01-21. 2024.

[214] Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, and Houston, Ann. *OntoNotes Release 5.0*. Philadelphia, 2013. DOI: 10.35111/xmhb-2b84. URL: https://catalog.ldc.upenn.edu/LDC2013T19.

[215] Wimalasuriya, Daya C. and Dou, Dejing. "Ontology-based information extraction: An introduction and a survey of current approaches". In: *J. Inf. Sci.* 36.3 (2010), pp. 306–323. ISSN: 0165-5515. DOI: 10.1177/0165551509360123. URL: https://doi.org/10.1177/0165551509360123.

[216] Wittgenstein, Ludwig. *Philosophical Investigations*. Basil Blackwell, 1953. ISBN: 0631119000.

[217] Xie, Xin, Zhang, Ningyu, Li, Zhoubo, Deng, Shumin, Chen, Hui, Xiong, Feiyu, Chen, Mosha, and Chen, Huajun. "From Discrimination to Generation: Knowledge Graph Completion with Generative Transformer". In: *Companion Proceedings of the Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing Machinery,

2022, pp. 162–165. ISBN: 9781450391306. DOI: 10.1145/3487553.3524238.

[218] Xilong, Hou, Junhan, Zang, and Xiaoguang, Wang. "Leveraging Large Language Models for Classification of Cultural Heritage Domain Terms: A Case Study on CIDOC CRM". In: *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*. Association for Computing Machinery, 2025. ISBN: 9798400710933. URL: https://doi.org/10.1145/3677389.3702562.

[219] Yang, Yiming. "An Evaluation of Statistical Approaches to Text Categorization". In: *Information Retrieval* 1.1-2 (1999), pp. 69–90. DOI: 10.1023/A:1009982220290.

[220] Yao, Yuan, Ye, Deming, Li, Peng, Han, Xu, Lin, Yankai, Liu, Zhenghao, Liu, Zhiyuan, Huang, Lixin, Zhou, Jie, and Sun, Maosong. "DocRED: A Large-Scale Document-Level Relation Extraction Dataset". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Korhonen, Anna, Traum, David, and Màrquez, Lluís. Florence, Italy: Association for Computational Linguistics, 2019, pp. 764–777. DOI: 10.18653/v1/P19-1074. URL: https://aclanthology.org/P19-1074/.

[221] Yates, Alexander, Banko, Michele, Broadhead, Matthew, Cafarella, Michael, Etzioni, Oren, and Soderland, Stephen. "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)*. Ed. by Carpenter, Bob, Stent, Amanda, and Williams, Jason D. Association for Computational Linguistics, 2007, pp. 25–26. URL: https://aclanthology.org/N07-4013/.

[222] Ye, Hongbin, Zhang, Ningyu, Chen, Hui, and Chen, Huajun. "Generative Knowledge Graph Construction: A Review". In: *Proceedings of the*

2022 Conference on Empirical Methods in Natural Language Processing. Ed. by Goldberg, Yoav, Kozareva, Zornitsa, and Zhang, Yue. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 1–17. DOI: 10.18653/v1/2022.emnlp-main.1.

[223] Yuan, H., Li, Y., Wang, B., and al., et. "Knowledge graph-based intelligent question answering system for ancient Chinese costume heritage". In: *npj Heritage Science* 13 (2025). Received 04 December 2024; Accepted 07 May 2025; Published 21 May 2025, p. 198. DOI: 10.1038/s40494-025-01776-x. URL: https://doi.org/10.1038/s40494-025-01776-x.

[224] Yuan, Weizhe, Neubig, Graham, and Liu, Pengfei. "BartScore: Evaluating generated text as text generation". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 27263–27277.

[225] Zaratiana, Urchade, Tomeh, Nadi, Holat, Pierre, and Charnois, Thierry. "GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Duh, Kevin, Gomez, Helena, and Bethard, Steven. Association for Computational Linguistics, 2024, pp. 5364–5376. DOI: 10.18653/v1/2024.naacl-long.300. URL: https://aclanthology.org/2024.naacl-long.300.

[226] Zhang, Wenxuan, Deng, Yue, Liu, Bing, Pan, Sinno, and Bing, Lidong. "Sentiment Analysis in the Era of Large Language Models: A Reality Check". In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Duh, Kevin, Gomez, Helena, and Bethard, Steven. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 3881–3906. DOI: 10.18653/v1/2024.findings-naacl.246. URL: https://aclanthology.org/2024.findings-naacl.246/.

[227] Zhang, You, Wang, Jin, Yu, Liang-Chih, and Zhang, Xuejie. "MA-BERT: Learning Representation by Incorporating Multi-Attribute Knowledge in Transformers". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Ed. by Zong, Chengqing, Xia, Fei, Li, Wenjie, and Navigli, Roberto. Online: Association for Computational Linguistics, 2021, pp. 2338–2343. DOI: 10.18653/v1/2021.findings-acl.206. URL: https://aclanthology.org/2021.findings-acl.206/.

[228] Zhao, Fudie. "A systematic review of Wikidata in Digital Humanities projects". In: *Digital Scholarship in the Humanities* 38.2 (2022), pp. 852–874. ISSN: 2055-7671. DOI: 10.1093/llc/fqac083. eprint: https://academic.oup.com/dsh/article-pdf/38/2/852/50488385/fqac083.pdf. URL: https://doi.org/10.1093/llc/fqac083.

[229] Zhong, Xiaoshi, Sun, Aixin, and Cambria, Erik. "Time Expression Analysis and Recognition Using Syntactic Token Types and General Heuristic Rules". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Barzilay, Regina and Kan, Min-Yen. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 420–429. DOI: 10.18653/v1/P17-1039. URL: https://aclanthology.org/P17-1039/.

[230] Zou, Qing and Park, Eun G. "Archival context, provenance, and a tool to capture archival context". en. In: *Arch. Sci.* 24.4 (2024), pp. 801–824.