

# Knowledge Graphs Generation from Cultural Heritage Texts: Combining LLMs and Ontological Engineering for Scholarly Debates

Andrea Schimmenti<sup>1</sup>, Valentina Pasqual<sup>1</sup>, Fabio Vitali<sup>1</sup>, Marieke van Erp<sup>2</sup>

<sup>1</sup>Università degli Studi di Bologna, Bologna, Italy  
{andrea.schimmenti2, valentina.pasqual2, fabio.vitali}@unibo.it

<sup>2</sup>KNAW Humanities Cluster, Amsterdam, the Netherlands  
marieke.van.erp@dh.huc.knaw.nl

## Abstract

Cultural Heritage texts contain rich knowledge that is difficult to query systematically due to the challenges of converting unstructured discourse into structured Knowledge Graphs (KGs). This paper introduces ATR4CH (Adaptive Text-to-RDF for Cultural Heritage), a systematic five-step methodology for Large Language Model-based Knowledge Extraction from Cultural Heritage documents. We validate the methodology through a case study on authenticity assessment debates.

**Methodology** - ATR4CH combines annotation models, ontological frameworks, and LLM-based extraction through iterative development: foundational analysis, annotation schema development, pipeline architecture, integration refinement, and comprehensive evaluation. We demonstrate the approach using Wikipedia articles about disputed items (documents, artifacts...), implementing a sequential pipeline with three LLMs (Claude Sonnet 3.7, Llama 3.3 70B, GPT-4o-mini).

**Findings** - The methodology successfully extracts complex Cultural Heritage knowledge: 0.96-0.99 F<sub>1</sub> for metadata extraction, 0.7-0.8 F<sub>1</sub> for entity recognition, 0.65-0.75 F<sub>1</sub> for hypothesis extraction, 0.95-0.97 for evidence extraction, and 0.62 G-EVAL for discourse representation. Smaller models performed competitively, enabling cost-effective deployment.

**Originality** - This is the first systematic methodology for coordinating LLM-based extraction with Cultural Heritage ontologies. ATR4CH provides a replicable framework adaptable across CH domains and institutional resources.

**Research Limitations** - The produced KG is limited to Wikipedia articles. While the results are encouraging, human oversight is necessary during post-processing.

**Practical Implications** - ATR4CH enables Cultural Heritage institutions to systematically convert textual knowledge into queryable KGs, supporting automated metadata enrichment and knowledge discovery.

**Keywords:** Digital Humanities; Cultural Heritage; Historical Documents; Scholarly Debate; Knowledge Representation; Ontology; Knowledge Extraction; Knowledge Graphs; Natural Language Processing; Large Language Models (LLM)

# 1 Introduction

Knowledge Graphs (KGs) have become the standard approach for representing and sharing Cultural Heritage (CH) information in the Linked Open Data (LOD) ecosystem, enabling interoperability between Libraries, Archives, and Museums institutions (Barabucci et al., 2021). This effort has been mostly concentrated on creating KGs of metadata, with diversified workflows dedicated to converting semi-structured or already structured sources (catalogues, inventories) into LOD (Bernasconi and Ferilli, 2024). However, the rich knowledge contained in unstructured CH texts — including descriptive content, contextual information, and analytical discourse — remains difficult to systematically extract and structure into queryable formats, and even when integrated into KGs, it is usually kept into long and description string fields (Barabucci et al., 2021; Giagnolini et al., 2025). Scholarly authenticity assessment debates exemplify this challenge, where complex interpretative knowledge is embedded in natural language discourse but practically absent from structured representations. Alongside the overarching challenges, additional ones stem from the inherently interpretative nature of humanities scholarship, which aligns with a constructivist epistemology viewing knowledge as situated, provisional, and shaped by the observer’s perspective. Checkland and Holwell (Checkland and Holwell, 2006) distinguish between *data*—passively recorded facts—and *capta*—knowledge actively constructed by the observer. This distinction challenges the realist assumption that often underpins data practices, in which data is treated as an objective and context-independent representation of reality.

These epistemological tensions manifest across various forms of CH scholarship, from attribution studies and provenance research to interpretative analysis and critical evaluation. Historical authenticity assessment exemplifies these challenges, as scholars from different humanities disciplines (e.g. Diplomatics, Palaeography, Philology, History) and scientific fields (e.g. Forensics, Materials science, Chemical analysis) frequently arrive at divergent conclusions based on different evidential priorities (Barone, 1912). Inherent factors contributing to this diversity include historical uncertainty, gaps in documentary transmission, and subjectivity (Blau, 2011; Gadamer, 2013).

Recent theoretical advancements acknowledge the subjectivity and uncertainty inherent in interpreting CH data, recognizing these as essential epistemic characteristics that must be preserved in digital representations (Pasqual, 2025; Piotrowski and Neuwirth, 2020; Piotrowski, 2023).

Despite these conceptual advances, current KG implementations represent only simplified versions of scholarly discourse. This representational gap between rich textual discourse and sparse structured data is systematic across CH domains. Whether dealing with artistic attribution, provenance disputes, historical interpretation, or authenticity assessment, complex scholarly reasoning gets reduced to simple categorical assertions. Major knowledge bases like Wikidata<sup>1</sup> and DBpedia<sup>2</sup> exemplify this limitation. While Wikipedia articles contain rich discussions with detailed scholarly arguments, evidence analysis, and alternative hypotheses, their structured counterparts reduce this complexity to sparse, categorical statements that fail to capture the evidential reasoning, methodological disagreements, and evolving consensus that characterize authentic scholarly discourse.

This pattern is systematic across CH materials: complex scholarly debates about document authenticity, artistic attribution, or historical interpretation are reduced to simple categorical assertions like “historical forgery” or “attributed to X,” stripping away the evidential reasoning, competing hypotheses, and methodological considerations that form the substance of scholarly discourse. While Wikidata employs a custom reification method to integrate claims with varying degrees of truthfulness through its ranking mechanism, annotators in the CH domain sometimes neglect this feature (Di Pasquale et al., 2024). Consider the so-called *Donation of Constantine*, a

---

<sup>1</sup><https://www.wikidata.org/>

<sup>2</sup><https://www.dbpedia.org/>

supposed 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 (Valla, 2023), demonstrating that its Latin contained anachronisms from the 8th rather than 4th century. Despite Valla’s compelling evidence, acceptance of this finding evolved gradually over centuries.

The screenshot shows the Wikidata item page for the 'Donation of Constantine'. At the top, there's a search bar and a navigation menu. Below the title 'Donation of Constantine (Q238476)', there are tabs for 'Item' and 'Discussion'. The main content area starts with a brief description: 'forged Roman imperial decree by which the emperor Constantine supposedly donated Rome and surrounding territory to the Pope Constitutum Constantini'. There are three main sections under 'Statements': 'instance of' (historical forgery), 'image' (a painting of a historical scene), and 'text' (a link to the DBpedia entry). Each statement has an 'edit' button and options to add references or values.

Figure 1: The Donation of Constantine entry in Wikidata

As shown in Figure 1, Wikidata categorizes the *Donation* as a "historical forgery"<sup>3</sup> with no representation of the scholarly debate, while DBpedia<sup>4</sup> similarly lacks structured representation of the authenticity discourse. In contrast, the corresponding Wikipedia page contains extensive discussions of Valla’s philological arguments, the specific linguistic evidence, the Church’s resistance, and subsequent scholarly confirmation. This fundamental **misalignment** between rich textual content and sparse structured claims illustrates the core challenge this research addresses.

This misalignment stems from two interconnected challenges. The first is *syntactic*: representing competing scholarly opinions within formal knowledge representation systems requires sophisticated mechanisms that traditional implementations struggle to handle effectively (Pasqual, 2025). While theoretical frameworks such as RDF-star, Named Graphs, and reification methods provide the necessary expressive power, their practical application demands complex modeling decisions about contradictory evidence, evolving consensus, and methodological disagreements, often resulting in oversimplified categorical assertions or unmanageably complex representations.

The second challenge is *practical*: extracting complex scholarly information from textual sources requires enormous manual labor, creating insurmountable scalability barriers. This process requires that expert annotators to identify scholarly agents, extract evidential reasoning, and capture alternative hypotheses while maintaining consistency across large document collections, a prohibitively expensive undertaking for most CH institutions that combines technical knowledge representation skills with deep humanities scholarship expertise.

CH institutions possess vast textual resources containing sophisticated scholarly analyses, but lack practical means to transform this knowledge into queryable, machine-readable formats. Large Language Models (LLMs) present a promising solution due to their sophisticated ability to parse complex academic discourse, identify implicit relationships, and handle domain-specific vocabularies

<sup>3</sup>Donation of Constantine - Q238476

<sup>4</sup>Donation of Constantine - DBpedia entry

(Khorashadizadeh et al., 2024). LLMs do not require large corpora of annotated data and can exploit transfer learning on different domains than those they were explicitly trained on (Brown et al., 2020a).

**Research Questions.** This work tackles the following primary research question: *How can a systematic methodology coordinate LLM-based Knowledge Extraction (KE) with ontological frameworks to effectively capture and structure the complex interpretative knowledge contained in CH texts?*

To systematically address this primary question, we investigate the following sub-questions, which we answer using our authenticity assessment case study:

1. **Methodological Framework:** What methodological approach can effectively coordinate LLM-based KE with existing ontological frameworks to capture complex scholarly interpretations in CH texts?
2. **Extraction Performance:** How accurately can systematic LLM-based pipelines extract different components of scholarly discourse, including metadata, agents, evidential reasoning, and interpretative hypotheses?
3. **Representation Fidelity:** Do automatically generated Knowledge Graphs adequately represent the complexity and nuance of scholarly interpretations when following structured methodological approaches?
4. **Model Comparison:** How do different LLMs perform within structured extraction pipelines for CH texts, and what are the implications for cost-effective deployment?
5. **Methodology Validation:** What insights does authenticity assessment validation provide about the methodology’s broader applicability to other forms of CH interpretative scholarship?

To address these research questions, we developed ATR4CH (Adaptive Text-to-RDF for Cultural Heritage), a systematic five-step methodology that combines annotation development, ontological alignment, and LLM-based extraction. We validate this methodology through authenticity assessment debates, a challenging domain that exemplifies the complex evidential reasoning, scholarly disagreement, and multi-perspectival structures characteristic of CH interpretative scholarship.

Our contributions are threefold. First, we present the ATR4CH methodology itself, providing a replicable framework for systematic LLM-based KE that can be adapted across CH domains and institutional resources. Second, we demonstrate the practical implementation by connecting ATR4CH to the SEBI ontology to develop annotation models and extraction pipelines for complex scholarly discourse. Third, we provide a comprehensive evaluation on a manually annotated sample of Wikipedia articles, establishing performance benchmarks across multiple extraction tasks and model architectures.

Our evaluation demonstrates the methodology’s effectiveness, achieving  $F_1$ -scores of 0.96-0.99 for metadata extraction, 0.7-0.8 for scholarly entity recognition, 0.65-0.75 for hypothesis extraction, and 0.95-0.97 for evidence extraction, with 0.62 G-EVAL overall discourse representativeness. Notably, smaller models performed competitively with larger architectures, indicating cost-effective deployment potential for resource-constrained CH institutions.

The remainder of the paper is organized as follows: Section 2 reviews related works in knowledge representation, extraction methods for the Semantic Web, opinion mining, and LLM-based KE approaches. Section 3 presents the ATR4CH methodology, detailing our five-step iterative approach for coordinating LLM-based extraction with ontological frameworks for the CH domain. Section 4

describes our corpus, describes the SEBI ontology, and explains the development of our annotation model through INCEPTION. Section 5 presents the implementation of LLM-based pipeline for our specific case. Section 6 provides detailed experimental results across five evaluation questions, comparing the performance of Claude Sonnet 3.7, Llama 3.3 70B, and GPT-4o-mini on metadata extraction, entity recognition, evidence mining, hypothesis extraction, and overall KE fidelity. Finally, Section 7 discusses our findings in relation to the presented research questions, analyzes performance trade-offs, addresses deployment implications, and outlines contributions, limitations, and future research directions. The methodology is designed to be adaptable across CH domains requiring extraction of multi-perspectival interpretative knowledge, with authenticity assessment serving as a validation case that demonstrates the approach’s effectiveness on particularly complex scholarly discourse. To ensure reproducibility of this work, all the code is published on a GitHub repository (<https://anonymous.4open.science/r/SEBI-Knowledge-Extraction-5FCE>).

## 2 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 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. KE is intimately connected to KR. One influences the other, and depending on the specific KR model adopted, KE practices must be adapted accordingly (Maynard et al., 2017).

### 2.1 Knowledge Representation

Recent theoretical advancements acknowledged the subjectivity and uncertainty inherent in interpreting CH data, recognizing these aspects as essential epistemic characteristics in analyzing and representing such data (Piotrowski and Neuwirth, 2020). Uncertainty in the CH domain arises not only from the data itself (for example, data extracted from the digitisation of a birth certificate) but also from the interpretative connections made by scholars regarding such data, such as identifying a name on a birth certificate with a specific historical figure (Piotrowski, 2023). However, these advancements have not translated into widely adopted practical tools and standards in KGs. Linked Open Data (LOD) is the standard for encoding and publishing CH data on the Web, promoting interoperability and data exchange between institutions. Standard online catalogues (e.g., Europeana)<sup>5</sup> typically provide single-perspective flat metadata, relegating discussions, debates, and uncertain facts to free text descriptions (Barabucci et al., 2021).

To the best of our knowledge, Wikidata is the only large-scale data catalogue 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 underutilise this feature. Additionally, claims related to CH data often make use of numerous qualifiers to encode contextual metadata, likely due to the increased effort required for this type of annotation (Di Pasquale et al., 2024).

Some ontologies have been designed to structure multi-perspective representations in CH data. ICON (Sartini et al., 2023; Baroncini et al., 2023) encodes visual recognitions in art history using  $n$ -ary relations to encode contextual metadata. Digital Hermeneutics (Daquino et al., 2020) employs a layered approach using Named Graphs (Carroll et al., 2005) to represent scholarly interpretations

---

<sup>5</sup><https://www.europeana.eu/>

in archival and literary sources. HiCo (Daquino and Tomasi, 2015) and the STAR model (Andrews, 2023) have been designed to represent historical interpretations and arguments. Wider adoption is hampered by the absence of tools to support the extraction, categorization, and contextualization of scholarly interpretations at scale, where practical effectiveness ultimately depends on the ability to extract such interpretations from unstructured sources. This is the gap the work presented in this paper aims to fill: it introduces a method and supporting infrastructure for the formalization and publication of interpretative claims within CH datasets, addressing the need for representations of scholarly discourse.

## 2.2 Knowledge Extraction for the Semantic Web

In the Semantic Web domain, KG generation can involve Graph Neural Networks (GNNs) capable of generating, enriching or correcting valid RDF from multiple sources. In the CH domain, KG generation is usually based on mappings from other sources, with some exceptions such as CIDOC2VEC (El-Hajj and Valleriani, 2021) building similarity recommendations using recurrent paths in CIDOC-CRM based KGs. Generally, generating KGs from text is closely linked to NLP tasks of Named Entity Recognition (NER), Relationship Extraction (RE), Entity Linking (EL) and similar subtasks (Maynard et al., 2017), commonly termed Text2KG.

Recent CH works demonstrate diverse Text2KG approaches. The Musical Meetups Knowledge Graph (MMKG) (Morales Tirado et al., 2023) leverages Wikipedia biographies and DBpedia Spotlight for entity recognition, using existing LOD as filters and enrichment sources. MusicBO (Gangemi et al., 2024) uses Text2AMR2FRED (Meloni et al., 2017), employing Abstract Meaning Representation (AMR) mapping to RDF/OWL with BLEURT score quality control. The Odeur-ropa project (Lisena et al., 2022) integrates annotation models with ontological representation through frame-based annotation schemes that map directly to CIDOC-CRM extended concepts, demonstrated through multilingual BERT-based models across seven European languages.

## 2.3 Opinion Mining for the Semantic Web

These approaches reveal three paradigms in CH Text2KG extraction: leveraging existing encyclopedic resources with focused ontologies (MMKG), employing sophisticated semantic parsing with quality validation (MusicBO), and designing integrated annotation-ontology frameworks (Odeur-ropa), underscoring the need for flexible methodologies addressing varying source types and resource constraints.

Aspect-Based Sentiment Analysis (ABSA) through SemEval 2014 (Pontiki et al., 2014) established granular opinion extraction frameworks with four subtasks: aspect term extraction, aspect term polarity, aspect category detection, and aspect category polarity. Recent advancements expanded ABSA to diverse datasets (Dong et al., 2014; Gao et al., 2019; Hamborg et al., 2021), with context-aware language models such as BERT improving performance (Gao et al., 2019) and LLMs such as GPT-3.5 achieving state-of-the-art results with zero-shot prompting (Wang et al., 2023). 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).

## 2.4 LLMs for Knowledge Extraction

LLMs have introduced new text-to-KG extraction paradigms (Khorashadizadeh et al., 2024; Mihindukulasooriya et al., 2023; Meyer et al., 2024). Allen et al. (Allen et al., 2023) identify two

primary directions: hybrid neuro-symbolic systems and natural language interfaces for domain experts—particularly relevant for CH contexts where experts lack technical expertise but possess deep interpretative knowledge.

Lairgi et al. (Lairgi et al., 2024) address traditional pipeline limitations through iText2KG’s zero-shot, incremental approach with four modules enabling dynamic knowledge base expansion. Kumar et al. (Kumar et al., 2022) demonstrate Knowledge Language Models (K-LM) injecting domain-specific RDF triples into language model architectures, revealing that relevance matters more than quantity of injected KGs. Ringwald (Ringwald, 2024) explores pattern-based extraction methods learning from Wikipedia-DBpedia/Wikidata pairs, offering middle ground between rigid rules and unpredictable LLM generation while maintaining alignment with established ontological frameworks.

### 3 The ATR4CH Methodology

This section introduces the *Adaptive Text-to-RDF for Cultural Heritage* (ATR4CH) methodology, an iterative approach specifically designed for extracting KGs from documents using LLMs in the CH, and by extension, Humanities, domain. Unlike traditional methodologies that treat annotation, KE, and ontology alignment as separate processes, ATR4CH recognizes that these processes are fundamentally interdependent in humanities contexts and must therefore be approached not sequentially but conjunctively.

#### 3.1 Methodology Overview

The ATR4CH methodology systematically transforms three foundational inputs into a validated KE system. The process begins with a corpus of unstructured documents, a target ontology defining the desired knowledge representation, and a set of Competency Questions (CQs) specifying what the system should be able to answer. Through five sequential steps, these inputs are transformed into a working extraction pipeline, a refined annotation model, and a comprehensive evaluation framework. A flowchart of the methodology is shown in Figure 2.

The methodology is designed to converge on a solution that meets both technical KE requirements and ontological representation standards for CH information.

Several foundational approaches in ontology engineering and KE are behind this methodology. The eXtreme Design methodology (Presutti et al., 2009) provides the theoretical foundation for iterative and incremental development in ontology engineering, emphasizing the centrality of CQs throughout the development process, an approach successfully applied in CH contexts such as Viewsari, a KG of Giorgio Vasari’s The Lives (Ondraszek et al., 2024). Additional inspirations include (Tomasi, 2020) for the selection of relevant items during the development phase, and the integrated annotation-ontology methodology used in Odeuropa’s annotation model for mentions of smell throughout CH documents (Lisena et al., 2022), which demonstrated how annotation schemas can be designed to map directly to ontological concepts within CH applications.

ATR4CH specifically focuses on coordinating an annotation and extraction pipeline with an ontology or combination of multiple ontologies in the CH domain, such as CIDOC-CRM (Doerr, 2003), Dublin Core<sup>6</sup>, FRBR/FRBRoo (IFLA Working Group on FRBR/CRM Dialogue, 2017), HiCo (Daquino and Tomasi, 2015), SKOS<sup>7</sup>, and PROV-O (Lebo et al., 2013). The methodology is suited on unstructured texts (informative, narrative, scholarly sources) and less for semi-structured

<sup>6</sup><https://www.dublincore.org/>

<sup>7</sup><https://www.w3.org/TR/skos-reference/>

documents such as catalogues or inventories, which are also common sources in the Libraries, Archives and Museum (LAM) domain. The ATR4CH methodology recognizes LLMs as particularly useful for their capabilities in performing KE through In-Context Learning (ICL) (Brown et al., 2020b), Few-Shot, and Chain-of-Thought (CoT) strategies (Petroni et al., 2019; Lairgi et al., 2024), based on established practices of KE (Tamasauskaitė and Groth, 2022).

ATR4CH assumes the existence of at least one document containing information that can be mapped to a given ontology, where this information must be annotated to answer a set of CQs. The approach is designed to be resource-adaptive, accommodating varying computational capabilities and project scales while maintaining consistent theoretical foundations.

### 3.2 Foundational Analysis and Design (Step I)

The first step establishes foundational understanding of the source material by analyzing both corpus and ontology to identify core patterns for KE. This analysis addresses potential data sparseness problems, which are common throughout Information Extraction tasks on unstructured texts.

- **Corpus Analysis:** This preliminary step examines how knowledge manifests throughout textual discourse, including linguistic patterns, discourse structures, and representational strategies. 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. In the case of Wikipedia articles about forged CH items, this analysis reveals specific patterns in both structure and content. Structurally, it identifies which sections contain scholarly opinions versus out-of-scope debates, enabling focused extraction from high-density sections like "Scholarly analysis" while filtering out biographical or tangential material. Content-wise, it determines whether articles present complete scholarly reasoning or merely final judgments (e.g., "most scholars consider this a forgery" versus detailed evidential arguments), guiding the methodology toward sources with sufficient depth for comprehensive KE.
- **Ontology Analysis:** This analysis 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 for extraction, which relationships can be reliably inferred from corpus discourse patterns, and which elements may need omission due to lack of textual support. Competency Questions guide prioritization of ontological coverage based on research requirements. For instance, in the case of the Donation of Constantine, a relevant CQ would be "What are the latest scholars identifying the document as authentic?"
- **Core Ontological Patterns (COPs) Identification:** Based on corpus and ontology analyses, this process identifies essential Knowledge Graph patterns required to answer the CQs. Core Ontological Patterns represent the central ontological nodes and relationships that are both present in the corpus as extractable information and necessary for addressing the research questions. For example, in authenticity assessment debates, a typical pattern would be "Scholar X evaluates Feature Y of Document Z using Method M and concludes Authenticity Status S." The identification process involves: (1) assessing alignment between Competency Questions, ontological structures, and available textual content, (2) identifying patterns with sufficient textual evidence for reliable extraction, (3) prioritizing based on extractability fea-

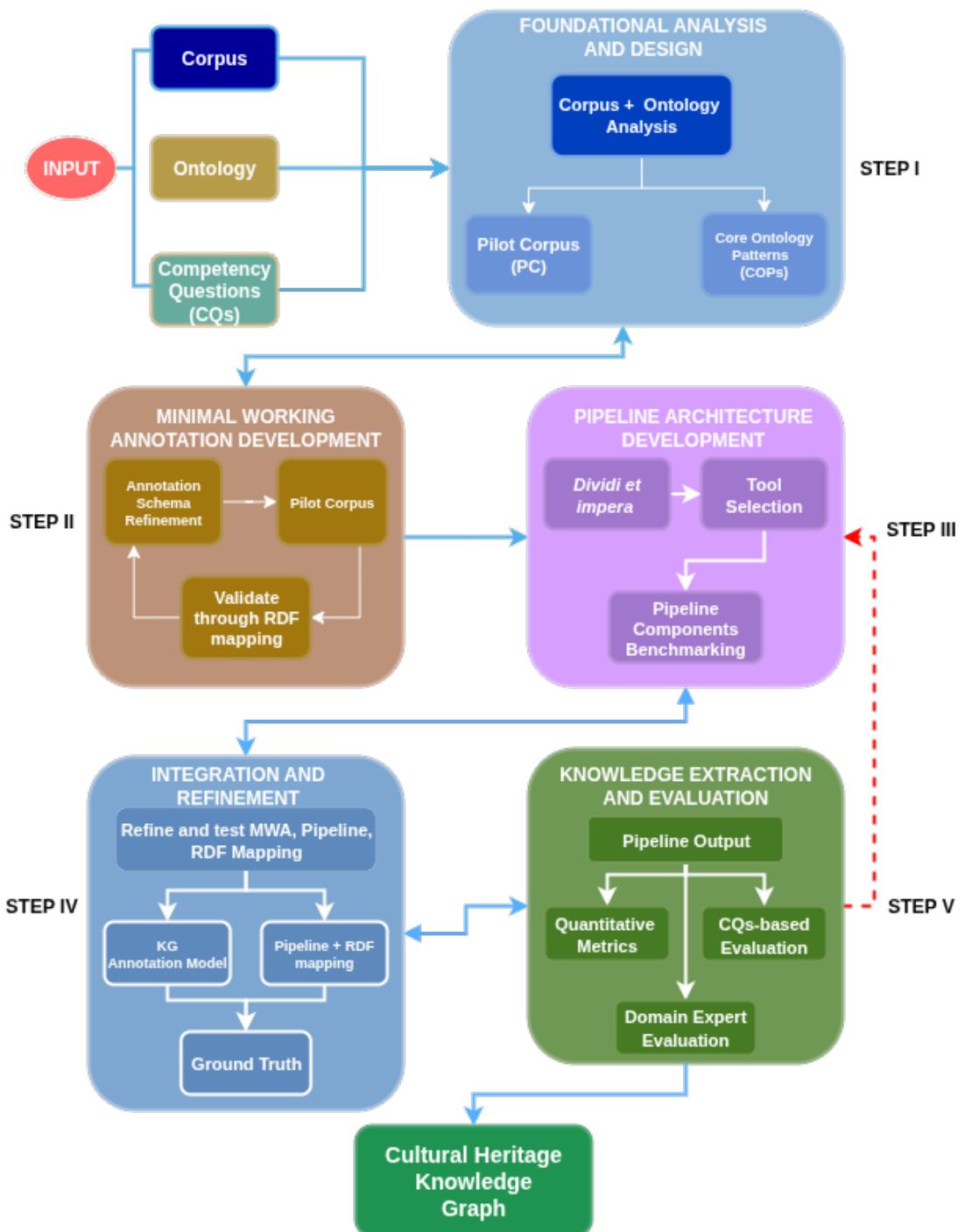


Figure 2: Flowchart of the ATR4CH methodology showing the five-step iterative process

sibility and CQ relevance, and (4) selecting a manageable subset that forms the semantic backbone for KE.

- **Pilot Corpus Selection:** The Pilot Corpus is a set of representative documents from the whole corpus that serves as a development sandbox for annotation and KE exploration. The pilot corpus is not intended as a quantitatively representative sample but rather as a qualitative sample that must be linguistically, structurally, and epistemically representative while remaining manageable for intensive manual development work. Selection criteria focus on ensuring coverage of various linguistic patterns, discourse structures, and diverse manifestations of the target COPs. The size can be as small as 3-5 documents, depending on document length and complexity of information manifestation patterns detected during corpus analysis.

### 3.3 Minimal Working Annotation Development (Step II)

The second step consists of iterative cycles to develop the Minimal Working Annotation (MWA). The aim of this step is to model, enrich, refine, and validate the MWA over the pilot corpus, and test whether the annotated data can be successfully mapped to KGs that satisfy the COPs identified in Step 3.2.

**MWA Schema Development:** Starting from the COPs identified in Step 3.2, this process develops an annotation schema that captures the essential knowledge structures while remaining practical for both manual annotation and automated extraction. The schema design must account for the diverse ways knowledge manifests in the corpus, as identified during the corpus analysis, including both explicit textual mentions and information that requires inference or contextualization.

The MWA should prioritize simplicity and feasibility while ensuring adequate coverage of the COPs. The "minimal" aspect refers to including only those annotation elements that are necessary for extracting the identified ontological patterns, avoiding over-annotation that may complicate the extraction process without contributing to answering the CQs. If the COPs require complex semantic structures beyond simple triple patterns, the annotation schema should include appropriate mechanisms for representing these relationships in a way that can be reliably mapped to RDF (e.g., if the ontology relies on Named Graphs, reification, etc).

**Knowledge Base Integration Strategy:** Knowledge base integration is an integral part of annotation processes because it enables consistent entity identification and vocabulary alignment between textual mentions and the target ontology. Since the COPs typically involve ontological individuals, entities, controlled vocabularies, or standardized terminologies defined within or referenced by the ontology, annotators need access to these resources to ensure that textual references are linked to the 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 in the annotation model development to establish clear protocols for entity linking and vocabulary alignment that will guide both manual annotation and automated extraction in Step 3.4. The choice between local and external knowledge bases depends on domain coverage, data quality requirements, and the specific entity types required by the COPs, with early integration ensuring that the annotation schema can consistently handle entity disambiguation, coreference resolution, and terminological standardization throughout the development process.

**Iterative Development Process:** The development follows a systematic cycle designed to ensure that the annotation schema can successfully produce RDF structures that satisfy the COPs:

1. **Schema Design:** Develop initial annotation layers based on the COPs, incorporating knowledge base integration protocols through tagsets, controlled vocabularies, and standardized terminologies that align with the target ontology.
2. **Pilot Corpus Annotation:** Annotate the entire pilot corpus using the current iteration of the MWA to identify potential gaps, inconsistencies, or practical bottlenecks in the annotation process.
3. **Mapping Validation:** Conduct preliminary mapping exercises from the annotated data to RDF format, testing whether the resulting KGs satisfy the COPs and adequately represent the semantic content of the source documents.
4. **Schema Refinement:** Refine the annotation model based on issues identified during mapping validation, returning to previous steps as necessary to address fundamental problems with the annotation approach.

These preliminary mapping exercises are crucial for validating that the annotation schema can produce the target knowledge structures. They serve as an early validation mechanism, ensuring that the annotation effort will ultimately generate RDF that satisfies the COPs before proceeding to automated extraction development.

### 3.4 Pipeline Architecture Development (Step III)

The third step designs and implements computational tools to automatically extract COPs from text using the MWA as target schema, addressing CH corpora's domain-specific characteristics and limited annotated training data.

**Task Decomposition and Architecture Design:** The KE task is designed around MWA elements, prioritizing based on COP semantic importance and accounting for information manifestation patterns from corpus analysis. This modular approach enables incremental extraction, where KG components are progressively identified through sequential processing, facilitating debugging and targeted optimization while minimizing error propagation through robust intermediate representations.

**Tool Selection Strategy:** Tool choice aligns with available resources and data characteristics:

- **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 use structured output generation through JSON schemas (Schick et al., 2023; Qin et al., 2024) and ICL strategies (Brown et al., 2020b; Min et al., 2022), combined with specialized NER tools (Devlin et al., 2019) for precise span identification.

**Pipeline Implementation:** Development targets the MWA schema, integrating knowledge base resources and vocabulary standardization protocols through prompt integration or RAG (Lewis et al., 2020). Initial implementation focuses on basic functionality across all COPs before optimization.

**Benchmarking Approach:** Evaluation strategies range from basic (standard metrics on pilot corpus) to comprehensive (ablation studies and hybrid approach exploration) based on project constraints.

**Output:** An extraction pipeline capable of processing raw text and generating structured outputs following the MWA schema.

### 3.5 Integration and Refinement (Step IV)

The fourth step harmonizes the COPs (Step 3.2), MWA (Step 3.3), and pipeline (Step 3.4) into a coherent, end-to-end KE system, bringing the experimental pipeline into production-ready status.

**End-to-End Pipeline Testing:** Comprehensive testing over the pilot corpus processes documents from raw text to final KGs, revealing systematic issues including data sparseness patterns, inconsistent tool coverage across discourse types, and representation generation errors. The testing systematically evaluates performance across document types and semantic phenomena, with particular attention to error propagation through pipeline stages.

**MWA Refinement to KG-AM:** Based on testing results, the MWA evolves into a production-ready KG Annotation Model (KG-AM) suitable for both manual annotation and automated extraction. This may involve adding elements crucial for automated extraction such as coreference chains, disambiguation tags, or confidence indicators while maintaining backward compatibility with COPs.

**Mapping Algorithm Enhancement:** Preliminary mapping algorithms from Step 3.3 are refined based on pipeline testing requirements, improving handling of complex semantic structures and adding validation using tools such as SHACL, OWL reasoners, SPARQLAnything (Asprino et al., 2023) and RML (Dimou et al., 2014). Error handling mechanisms manage extraction failures and partial results.

### 3.6 Knowledge Extraction and Evaluation (Step V)

The final validation phase employs technical validation and domain-expert evaluation to ensure knowledge structures accurately represent domain-specific discourse complexity, applying the refined system from Step 3.5 to test data.

**Ground Truth Preparation:** Comprehensive Ground Truth (GT) creation using the KG-AM involves annotating a test dataset separate from the pilot corpus, covering all COPs from Step 3.2 and applying mapping algorithms from Step 3.5 to generate reference RDF.

**Knowledge Extraction:** Test datasets are processed through the complete pipeline under realistic deployment conditions, with systematic documentation of performance and failure modes.

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

- **Technical Evaluation:** Component-level assessment using precision, recall, and  $F_1$ -score, plus coverage analysis for Competency Questions
- **Semantic Evaluation:** KG "rehydration" (Gardent et al., 2017; Gangemi et al., 2024) enables comparison when structural alignment is impossible, using metrics like BLEU (Pa-

pineni et al., 2002), METEOR (Banerjee and Lavie, 2005), BARTScore (Yuan et al., 2021), CHRF+++ (Popović, 2015), and G-EVAL (Liu et al., 2023) as proposed by (He et al., 2025)

- **Competency-Based Evaluation:** SPARQL query suites derived from original CQs, aligned with (Presutti et al., 2009) evaluation using tools like TestaLOD (Carriero et al., 2019)

**Domain Expert Validation:** Comprehensive review by domain specialists evaluates extraction quality and coherence, with rehydration technique enabling evaluation by experts without RDF expertise.

**Iteration Strategy:** Evaluation results may trigger returns to earlier steps: coverage issues to Step 3.3 or Step 3.5, extraction bottlenecks to Step 3.4, or systematic errors requiring architectural restructuring in Step 3.5.

## 4 Corpus, Ontology and Annotation Model

As stated in 3, our methodology starts from a corpus, an ontology, and a set of CQs. This section discusses how the dataset was collected and which documents it contains (4.1); the ontology (see 4.2); and the implementation of the first two steps of the methodology, namely Step I (see 3.2), and Step II (see 3.3).

### 4.1 Dataset

Our dataset comprises Wikipedia articles focusing on historical forgeries, hoaxes, and authenticity controversies across CH. The collection was compiled using automated web scraping from Wikipedia’s categorical organization system, targeting categories related to forgeries and authenticity debates.

We scraped articles from 15 distinct Wikipedia categories, retrieving full article text and inter-article link structures (sitelinks<sup>8</sup>). The script selects both categorical pages<sup>9</sup> and standalone articles<sup>10</sup>, storing each document with complete textual content and associated metadata including categorization and cross-references to related entities.

The initial selection covered 31 categories, spanning from Document<sup>11</sup> and Literary Forgeries<sup>12</sup> to Historical Myths<sup>13</sup>, Conspiracy Theories<sup>14</sup>, Pseudepigraphy (i.e. falsely attributed works, texts whose claimed author is not the true author, or works whose real author attributed it to a figure of the past)<sup>15</sup> and Political forgery<sup>16</sup>. Out of the total 1301 documents which were retrieved,<sup>17</sup>, 16 categories and 717 articles were excluded, as they did not present any scholarly debate or they were not about a CH item (be it a document, an artifact, etc). The dataset encompasses 581 articles as shown in Table 1.

<sup>8</sup><https://www.wikidata.org/wiki/Help:Sitelinks>

<sup>9</sup>See for instance the Wikipedia Category "Forgery" (<https://en.wikipedia.org/wiki/Category:Forgery>)

<sup>10</sup>See for instance the article describing the Donation of Constantine ([https://en.wikipedia.org/wiki/Donation\\_of\\_Constantine](https://en.wikipedia.org/wiki/Donation_of_Constantine))

<sup>11</sup>[https://en.wikipedia.org/wiki/Category:Document\\_forgeries](https://en.wikipedia.org/wiki/Category:Document_forgeries)

<sup>12</sup>[https://en.wikipedia.org/wiki/Category:Literary\\_forgeries](https://en.wikipedia.org/wiki/Category:Literary_forgeries)

<sup>13</sup>[https://en.wikipedia.org/wiki/Category:Historical\\_myths](https://en.wikipedia.org/wiki/Category:Historical_myths)

<sup>14</sup>[https://en.wikipedia.org/wiki/Category:Conspiracy\\_theories](https://en.wikipedia.org/wiki/Category:Conspiracy_theories)

<sup>15</sup><https://en.wikipedia.org/wiki/Category:Pseudepigraphy>

<sup>16</sup>[https://en.wikipedia.org/wiki/Category:Political\\_forgery](https://en.wikipedia.org/wiki/Category:Political_forgery)

<sup>17</sup>The selection was performed on October 2024

Table 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
<b>Total</b>	<b>581</b>

The corpus exhibits significant 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 the scholarly discourse on authenticity assessment. As shown in Figure 3, the distribution of article lengths follows a right-skewed pattern, with most articles ranging from 2k to 15k characters, with some outliers that extend beyond 40k characters.

The distribution across categories reflects the natural prevalence of different forgery types in scholarly discourse (Figure 4). Literary forgeries represent the largest category with 138 articles, followed by various forms of pseudepigraphy totaling 136 articles across subcategories. Archaeological and artistic forgeries comprise 132 articles combined, while more specialized categories such as musical hoaxes and religious controversies contain fewer but often more detailed entries.

The temporal scope of the corpus spans from late antiquity to the contemporary period, representing diverse scholarly debates across different historical contexts.

Notable variations exist across categories in terms of content density and scope, as illustrated in Figure 5. The token distribution reveals substantial variability within categories, with numerous outliers indicating comprehensive case studies that warrant extensive coverage. The Shakespeare authorship question category demonstrates the highest token density, with articles reaching nearly 10K tokens, reflecting the extensive scholarly debate surrounding this topic. Political forgery and religious hoaxes also show elevated token counts, indicating rich semantic content suitable for KE tasks. Conversely, categories like musical hoaxes and modern pseudepigrapha exhibit more consistent, moderate-length articles with fewer outliers.

Among the corpus, one of the oldest CH items under discussion is the Demodocus<sup>18</sup>, a fabricated Platonic dialogue that exemplifies early pseudepigraphic practices. For this specific case, a Wikidata entry exists<sup>19</sup> which appropriately employs a deprecated rank for the authorship claim linking Plato

<sup>18</sup>[https://en.wikipedia.org/wiki/Demodocus\\_\(dialogue\)](https://en.wikipedia.org/wiki/Demodocus_(dialogue))

<sup>19</sup><https://www.wikidata.org/wiki/Q2625856>

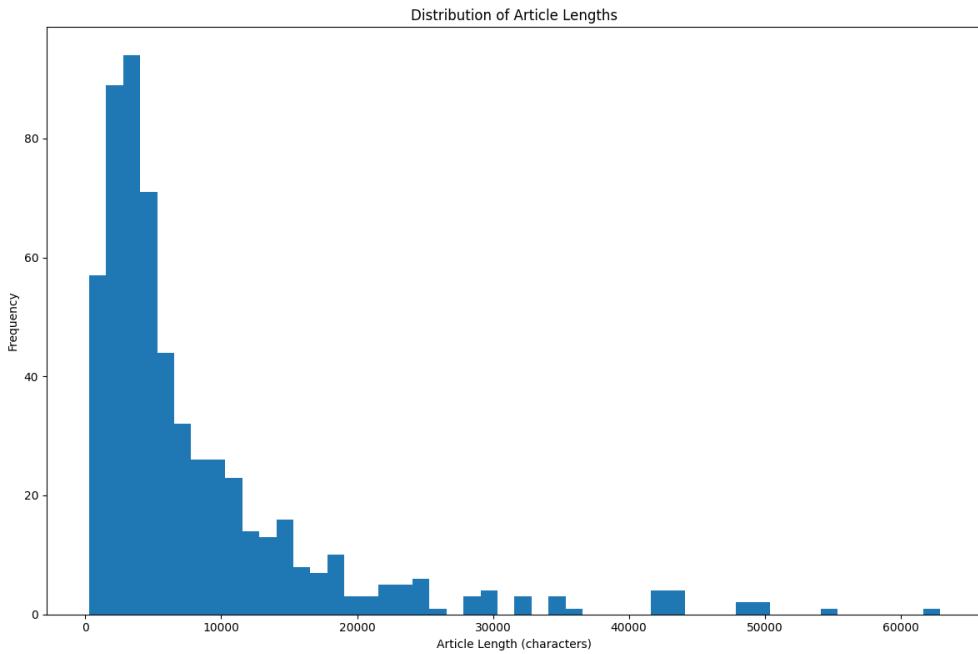


Figure 3: Overall distribution of article lengths showing the right-skewed pattern characteristic of encyclopedic content, with most articles in the 2k-15k character range and notable outliers extending beyond 40k characters.

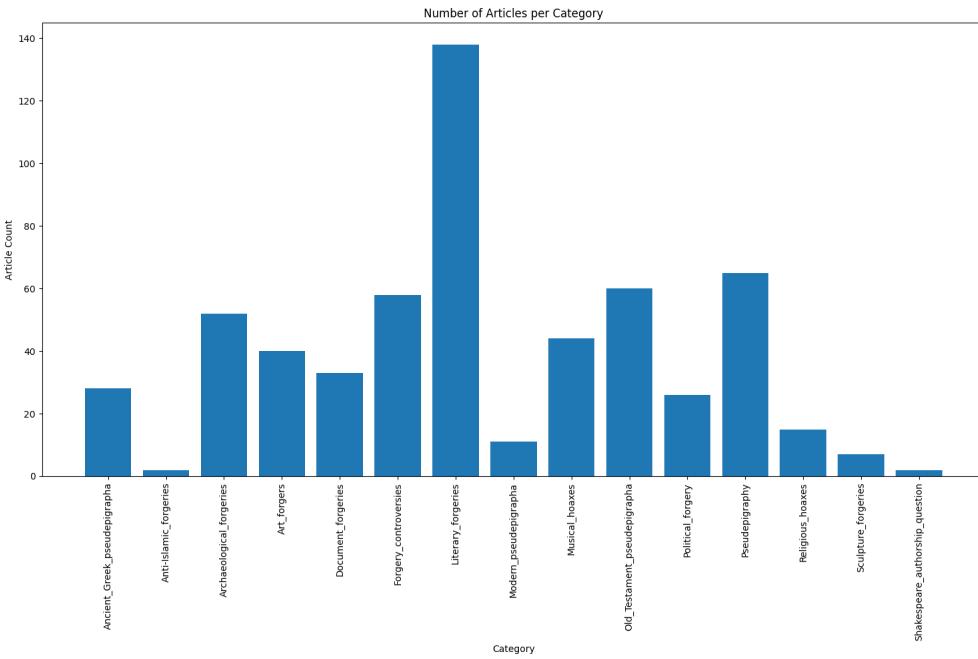


Figure 4: Distribution of articles across Wikipedia categories, showing the natural prevalence of different forgery types in scholarly discourse.

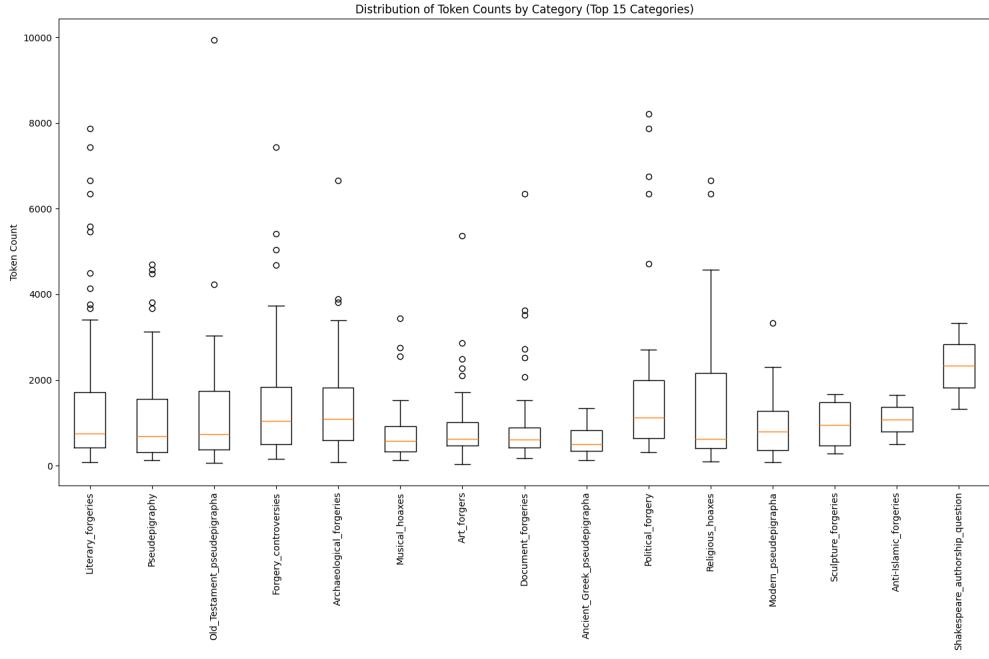


Figure 5: Token count distribution by category, illustrating variability in article length and content density. Box plots show medians, quartiles, and outliers representing comprehensive case studies.

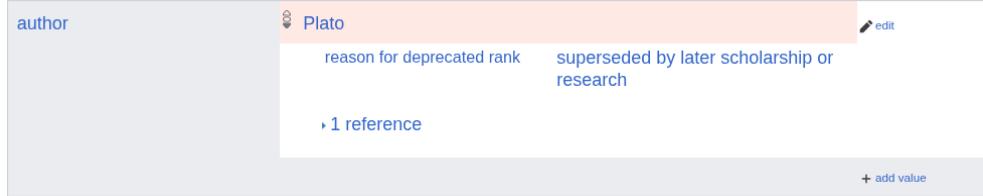


Figure 6: Plato noted as the author of the Demodocus using a deprecated rank, illustrating how existing knowledge bases can represent disputed attributions

to the Demodocus, as shown in Figure 6, as exemplified in Section 2.1.

Another article of the corpus is the Protocols of the Elders of Zion, one of the most famous forgeries to date<sup>20</sup>. Among the most recent examples, the 1996 *Posthumous Diary*<sup>21</sup>, an allegedly forged collection of poems by Italian poet Eugenio Montale, which caused a participated debate in the Italian philology community.

## 4.2 Ontology

The Scholarly Evidence Based Interpretation ontology<sup>22</sup>) (SEBI) (Pasqual, 2025) was based on a selection of scholarly articles, e.g. (Härtel, 2017), a catalogue describing 153 known forgeries from Styria (Haider, 2022), and several discussions with an expert Diplomatist. The data model represents authenticity assessment claims using RDF-star (Hartig, 2017) as a reification method to

<sup>20</sup>[https://en.wikipedia.org/wiki/The\\_Protocols\\_of\\_the\\_Elders\\_of\\_Zion](https://en.wikipedia.org/wiki/The_Protocols_of_the_Elders_of_Zion)

<sup>21</sup>[https://en.wikipedia.org/wiki/Posthumous\\_Diary](https://en.wikipedia.org/wiki/Posthumous_Diary)

<sup>22</sup><https://valentinapasqual.github.io/sebi/>

represent all (possibly concurrent) claim contents as well as their contextual information (Daquino et al., 2020). The content of each claim provides the following basic information about the document: authenticity classification, date and place of creation, author, and the intention behind its creation. Additionally, contextual information about the claim is recorded, specifically the evidences collected by the scholar to reach a certain conclusion (using HiCo),<sup>23</sup> evidence-based assessments as well as the author of the claim and relevant bibliographic entries (using PROV-o).<sup>24</sup> RDF-star (Hartig, 2017) has been chosen as the reification method to express both the claim contents and context, allowing the representation of the entire evaluation process conducted by scholars.

As shown in Figure 7 each claim contains an attempt of classification towards the CH item authenticity. This is obtained by making sure that all items are 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 `sebi:Document - dct:creator - dct:Agent`), the date of creation (`sebi:Document - dct:date - time:Interval`), location of creation (`sebi:Document - dct:coverage - dct:Location`) and the intention behind the document creation (`sebi:Document sebi:intended sebi:Intention`). The `dct:date` property is connected to a `time:Interval` class, which includes `time:hasBeginning` and `time:hasEnd` properties to specify the creation period and handle fuzzy time-spans.

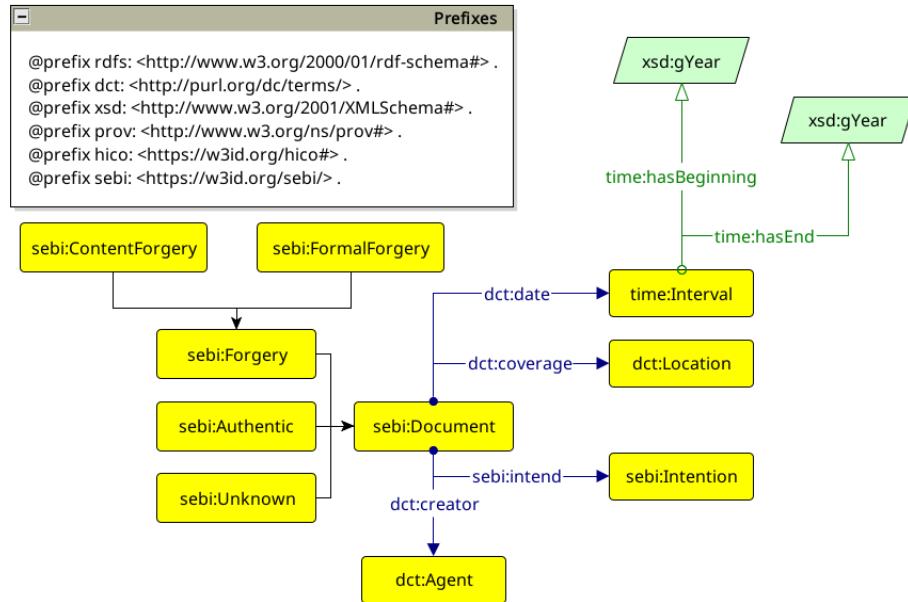


Figure 7: Selection of classes and properties to represent scholarly claims tackling authenticity assessment of a document

Concerning contextual information, each interpretation (set of claims represented as quoted triples) is categorised as a `hico:InterpretationAct` connected to a `prov:Agent` to address its auhtority and linked to the evidence supporting the claim (`sebi:support sebi:Evidence`). Document features and their evaluation are critical components of the ontology. Document features

<sup>23</sup><https://marilenadaquino.github.io/hico/>

<sup>24</sup><https://www.w3.org/TR/prov-o/>

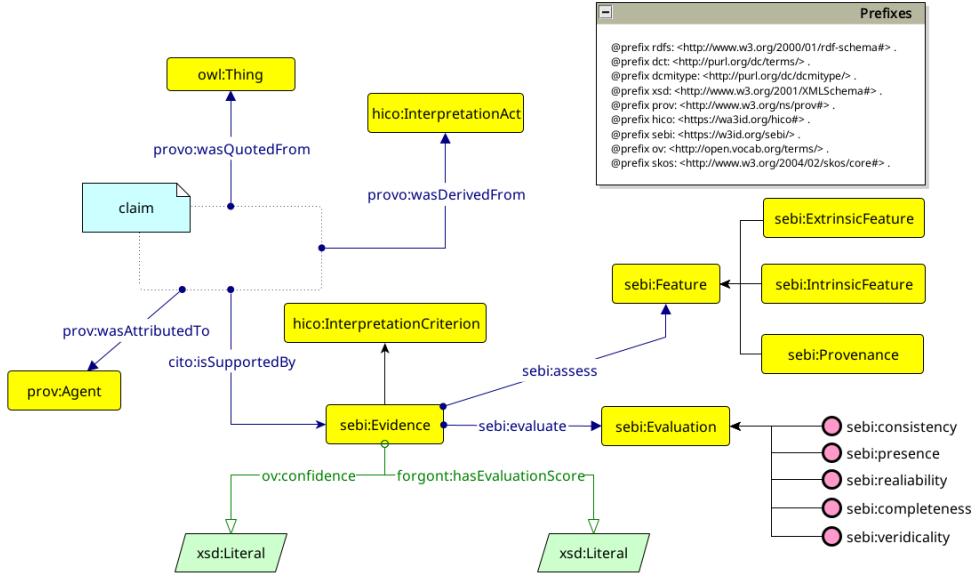


Figure 8: Selection of classes and properties to represent the contextual information about scholarly claim tackling authenticity assessment of a document

(**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 a set of established criteria (**sebi:Evaluation**) such as consistency, presence, completeness, veridicality, and reliability. A score is associated to each evidence as **xsd:Literal** using the property **forgont:hasEvaluationScore**. The evaluation score indicates a measure on each collected evidence, allowing the integration of negatives (e.g. the absence of the signature in a document is represented as an evidence based on the feature "authentication marks" with evaluation "presence", with score **false** or 0).

### 4.3 Pilot Corpus

Following Step II (3.2), we sampled the pilot corpus. We selected the following articles (*Donation of Constantine*, *Eremine Letter*, *Getty Kouros*, *Historia Augusta*, *Life of Homer*, *Marriage Charter of Empress Theophanu*, *Protocols of the Elders of Sion*), each belonging to a different category. The domain expert personally pointed to the *Donation of Constantine* and other medieval charters, as the *Marriage Charter*, as optimal use cases for the study.

The other articles were selected as they contained explicit scholarly disagreements about authenticity. Selection criteria included: (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.

### 4.4 Minimal Working Annotation

Our annotation model was developed through INCEpTION (Klie et al., 2018), following Step II of the methodology (3.3). The annotation model implements three core patterns from our ontology: **CH item metadata**, **scholarly agents**, and **authenticity opinions**.

**CH Item Metadata.** We established an identification layer 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`, `dcmitype:Collection`) with appropriate subclass relationships.

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

**Authenticity Claims.** We modeled claims 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`.

Using this approach on the Pilot Corpus, we validated annotation-to-RDF mapping through the algorithm in Listing 1, then enriched the model with additional ontology patterns.

Listing 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))

STEP 2: Extract CH Items
Select all spans marked as Entity
WHERE span has ItemTitle label
=> ItemSet(ItemSpan, WikidataID)

STEP 3: Find Relations
For CognizerSpan in CognizerSet, check if CognizerSpan
has stm:Object relation to span in ItemSet
=> Valid tuples (Cognizer, Item, Opinion)

STEP 4: Generate RDF for each tuple
For each matching pattern:
| -- Generate URI for Cognizer
| -- Add owl:sameAs + Wikidata ID
| -- Generate URI for Item
| -- Add owl:sameAs + Wikidata ID
| -- Map opinion to corresponding SEBI class (e.g., sebi:Forgery)
| -- Generate URI for Named Graph (hico:InterpretationAct)
| -- Generate claim triple as a RDF-star statement

+-- Apply template:

ex:{cognizer_uri}_about_{item_uri} rdf:type hico:InterpretationAct ;
prov:wasAttributedTo ex:cognizer .

ex:cognizer rdf:type dct:Agent ;
rdfs:label "CognizerSpan"@language .
owl:sameAs wd:wikidataId

ex:item rdf:type ex:type ;

```

```

rdfs:label "ItemSpan"@language .
owl:sameAs wd:wikidataID

<< ex:item rdf:type sebi:Opinion >> prov:wasDerivedFrom ex:{cognizer_uri}_about_{item_uri} .

```

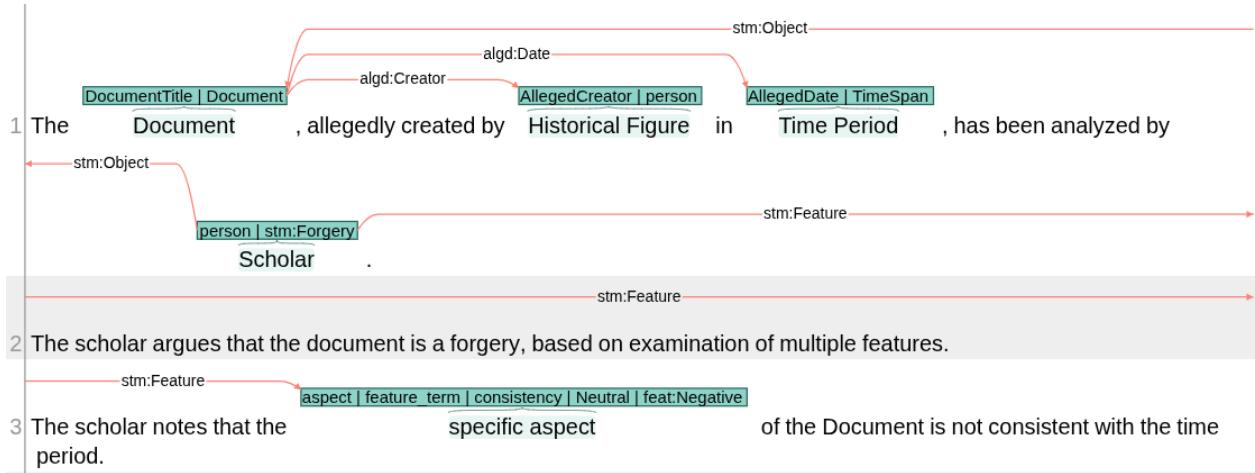


Figure 9: Example annotation of an entity expressing an opinion about a CH item

## 4.5 SEBI Annotation Model

The refined annotation model captures CH item metadata, evidence and features, and scholarly hypotheses through three additional layers, satisfying Step IV of the methodology (see 3.5).

**CH Item Metadata Layer.** This layer captures *alleged metadata*—the descriptive information (creator, date, location) that the document or artifact claims about itself, representing what the item purports to be before any scholarly critical analysis. This includes the face-value claims presented by the item regarding its 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).

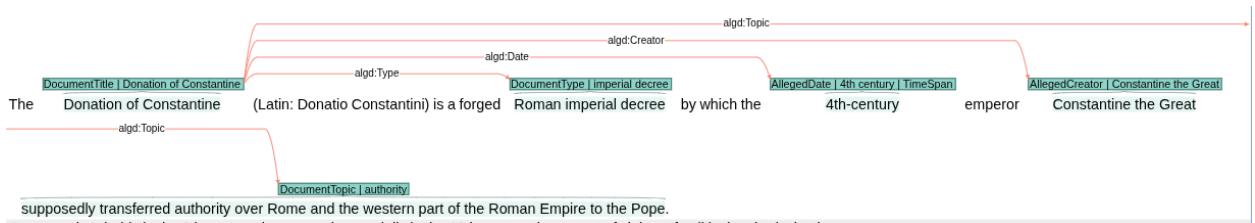


Figure 10: Alleged metadata annotation for the Donation of Constantine

**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, which converts to three evidence structures linking specific textual features to evaluation criteria and polarities.

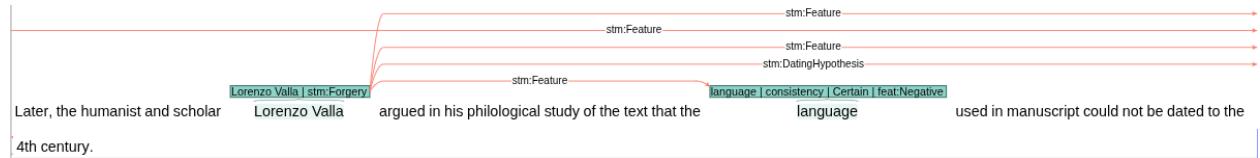


Figure 11: Lorenzo Valla's opinion with feature assessment annotation

Listing 2 shows the evidence mapping algorithm.

#### Listing 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**.

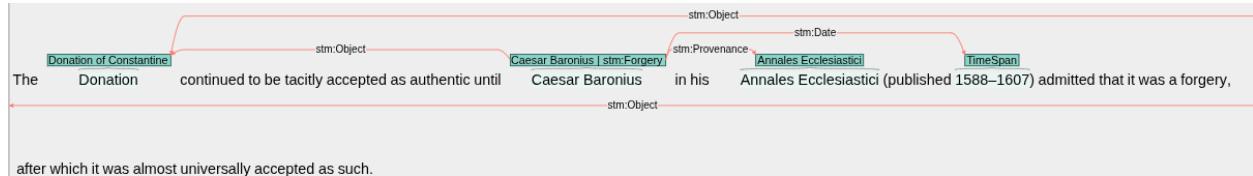


Figure 12: Caesar Baronius's admission of forgery with provenance annotation

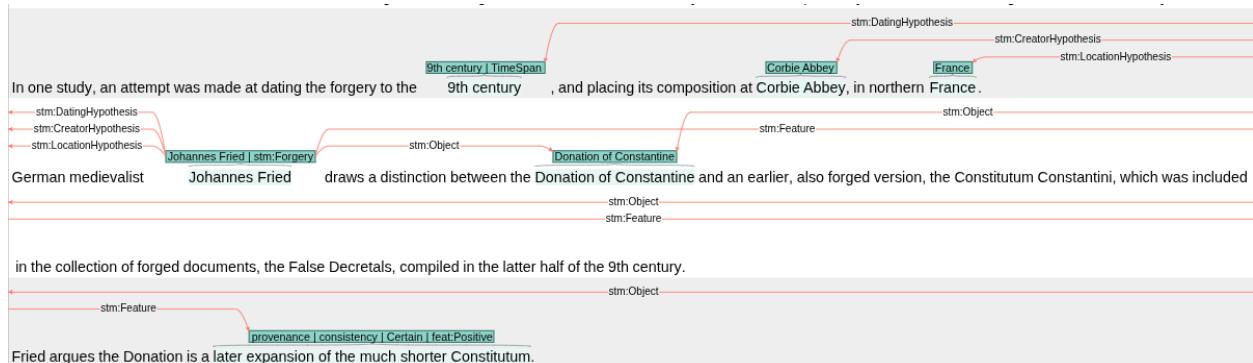


Figure 13: Johannes Fried's hypotheses annotation for the Donation of Constantine

Listing 3 details the hypotheses mapping algorithm.

### Listing 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)

STEP 3: Find Valid Patterns
For each relation in HypothesesSet:
Check if CognizerSpan exists in CognizerSet
=> Valid tuples (Cognizer, HypothesisType, Target)

```

```

STEP 4: Generate Target URIs
For each matching pattern:
| -- Generate/Reuse URI for Cognizer
| -- Generate/Reuse URI for Item
| -- Generate/Reuse URI for Target entity
| -- Map HypothesisType to corresponding RDF property

STEP 5: Generate RDF-star Statements

+-- Apply template:

kb:{target_uri} a {target_class} ;
    owl:sameAs wd:{wikidata_id} ;
    # if Wikidata ID not available
    # kb:{urifiedTargetSpan} a {target_uri} ;
    rdfs:label "{TargetSpan}"@language .

<< kb:{item_uri} dct:creator kb:{target_uri} >>
    prov:wasDerivedFrom kb:{cognizer_uri}_about_{item_uri} .

<< kb:{item_uri} dct:date kb:{target_uri} >>
    prov:wasDerivedFrom kb:{cognizer_uri}_about_{item_uri} .

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

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

```

Each annotation layer maps to RDF following SEBI ontology principles, with Wikidata integration providing entity resolution and the INCEpTION project available on GitHub alongside mapping scripts<sup>25</sup>. Statistics for the annotation results are shown in Table 2.

Table 2: Ground Truth Annotation results

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

## 5 Knowledge Extraction Pipeline and Evaluation Framework

This section presents the implementation of Steps III-V of our methodology (3.4, 3.5, 3.6), transforming the annotation model into a working KE pipeline. Our implementation integrates three complementary technologies in a sequential LLM-based process:

---

<sup>25</sup>SEBI-KE repository. See 'Inception2Graph' folder

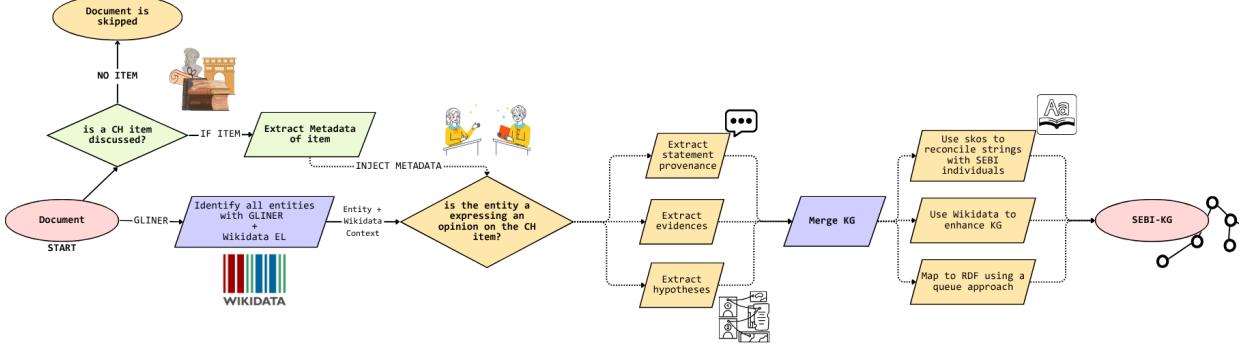


Figure 14: Flowchart of the sequential pipeline for SEBI-based KG generation

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

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** (Zaraticana et al., 2024) 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 three models with different parameter scales to understand performance trade-offs: Claude Sonnet 3.7 as the biggest model<sup>26</sup>, Llama 3.3 70B (Dubey et al., 2024) as a medium-sized model, and GPT-4o-mini.<sup>27</sup> While the exact parameter size of GPT-4o-mini remains undisclosed, estimates range from 8-14 billion active parameters (Ben Abacha et al., 2025). Figure 14 presents a comprehensive overview of the sequential processing pipeline.

**Entity Linking** employs a rule-based approach leveraging the Wikibase API<sup>28</sup> 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 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.

<sup>26</sup>Claude Sonnet 3.7 Model Card: <https://assets.anthropic.com/m/785e231869ea8b3b/original/clause-3-7-sonnet-system-card.pdf>

<sup>27</sup>GPT-4o-mini model card: <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>

<sup>28</sup>Wikibase API Documentation

## 5.1 Sequential Processing Pipeline

The KE pipeline is based on six components, each enriching the output before passing it to the next. Each component produces a JSON, each with a given schema designed to be convertible to RDF at the end. The development of the pipeline started from previous implementations and sequential testing over the COPs until the full KG could be extracted. As the output of the JSON model is largely similar to the output of the GT, the mapping works using a very similar logic, with the only difference of being mapped from JSON instead of 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.

### 5.1.1 CH item Metadata Extraction

**Aim:** Identify all the CH items being discussed in an article and get the alleged metadata about them

**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*, or, in other words, what items claim to be—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 15 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	JSON output
<p>The <b>Donation of Constantine</b> [...] 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.</p>	<pre>{   "item": "Donation of Constantine",   "alleged_author": "Constantine the Great",   "alleged_date": "4th century",   "alleged_location": "Rome",   "item_type": "decree", }</pre>

Listing 15: CH Item metadata extraction from source text to structured JSON output

### 5.1.2 Cognizer Identification

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

**Input:** Cleaned article + item(s) metadata (output of previous step, see Section 5.1.1)

**Output:** JSON with `is_cognizer` classification, coreferences

This component employs GliNER (Zaratiana et al., 2024) 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 selecting exclusively the subset of entities who actually 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 (5.1.1). 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 In-Context Learning (ICL) in a Few-Shot setting (with 3 examples), using Chain-of-Thought (COT) reasoning. Listing 16 shows a paragraph where **Lorenzo Valla** is mentioned alongside the output.

Input: JSON output	
<b>Input: Source text</b>  Later, the humanist and scholar <b>Lorenzo Valla</b> 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. <b>Valla</b> believed the forgery to be so obvious that he suspected that the Church knew the document to be inauthentic.	{ "entity": "Lorenzo Valla", "start": 605, "end": 618, "label": "person", "is_cognizer": true, "is_subject": true, "mentions": [ "Lorenzo Valla", "Valla" ] }

Listing 16: CH Item extraction from source text to structured JSON output

### 5.1.3 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<sup>29</sup> to retrieve candidates for the longest mention of each entity. Each candidate receives a score based on name similarity (Levenshtein distance<sup>30</sup>) between mentions and Wikidata labels/aliases, entity type compatibility, and for people, occupation relevance using Wikidata property P106<sup>31</sup> (prioritizing schol-

<sup>29</sup><https://www.mediawiki.org/wiki/Wikibase/API>

<sup>30</sup><https://pypi.org/project/python-Levenshtein/>

<sup>31</sup><https://www.wikidata.org/wiki/Property:P106>

arly occupations likely to express opinions in this domain).

Input: Clustered entities	Output: Wikidata entity
<pre>{   "primary_mention": "Lorenzo Valla",   "all_mentions": [     "Lorenzo Valla",     "Valla",     "the humanist"   ],   "entity_type": "person",   "paragraphs": [0, 3, 7] }</pre>	<pre>{   "wikidata_label": "Lorenzo Valla",   "wikidata_id": "Q214115",   "occupation": ["humanist",     "philologist"...],   "birth_year": 1407,   "death_year": 1457,   "mentions": ["Lorenzo Valla",     "Valla"] }</pre>

Listing 17: Entity resolution and linking: from clustered mentions to Wikidata-enriched entities

#### 5.1.4 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, the metadata of the opinion (where, when, provenance)

The fourth component extracts and classifies authenticity opinions based on 5.1.2. If the entity has been successfully linked to Wikidata, this information is given to the model as well.

The extraction process captures the main COP 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 18 as reference.

Input: Source text	JSON output: Opinion classification
<p>Later, the humanist and scholar <b>Lorenzo Valla</b> argued in his philological study of the text that the language used in <b>the manuscript</b> 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 <b>forgery</b> to be so obvious that he suspected that the Church knew the document [...]</p>	<pre>"opinions": [   {     "entity": "Lorenzo Valla",     "subject": "Donation of Constantine",     "opinion": "Forgery",     "confidence": "High",     "date": "1439-1440",     "location": ""   } ]</pre>

Listing 18: JSON output from source text about Cognizer, subject of the opinion, provenance, assessment

### 5.1.5 Evidence Mining and Feature Assessment

**Aim:** Enrich the Cognizer's opinion with the supporting evidence

**Input:** Structured opinions + contextual paragraphs

**Output:** JSON with supporting evidences and evaluations for each opinion

In the fifth component, the model is tasked with enriching 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/documentated?), 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 19 as reference.

Input: Source text	JSON output: Evidence extraction
[...] <b>Reginald Pecocke</b> , 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 <b>imperial-era formulas</b> are <b>used in the text</b> , some of the Latin in the document could not have been written in the 4th century	<pre>"evidence_evaluations": [   {     "evidence": "imperial-era formulas",     "feature": "language",     "evaluation": "presence",     "polarity": "positive"   } ]</pre>

Listing 19: JSON output of the identified evidences with evaluation

### 5.1.6 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 of the previous one 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 20 as 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 20: Hypothesis extraction from source text to structured alternative theories

## 5.2 Knowledge Graph

The final output is then mapped to RDF using the algorithms explained in Sections 4.4 and 4.5. This subsection serves as a showcase of the produced KGs in RDF-star and to describe the anatomy of our outputs (specifically, this was generated by the pipeline using Llama 3.3 70B). Figure 21 shows the general structure of a generated KG from the GraphDB interface (Ontotext, 2024). Each CH item is represented with both alleged metadata (what the item(s) claim(s) to be) and scholarly assessments, as shown in Listing 4. The Donation of Constantine exemplifies this pattern:

Listing 4: Document representation with alleged and scholarly metadata

```
# Basic Item information
kb:donation_of_constantine a sebi:Decree ;
  dct:title "Donation\u2022of\u2022Constantine"@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:
  march_30_no_year_specified_but_implied_to_be_during_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 5 shows Lorenzo Valla's interpretation of the Donation.

Listing 5: Lorenzo Valla's interpretation with supporting evidence

```
# Lorenzo Valla as scholarly agent
```

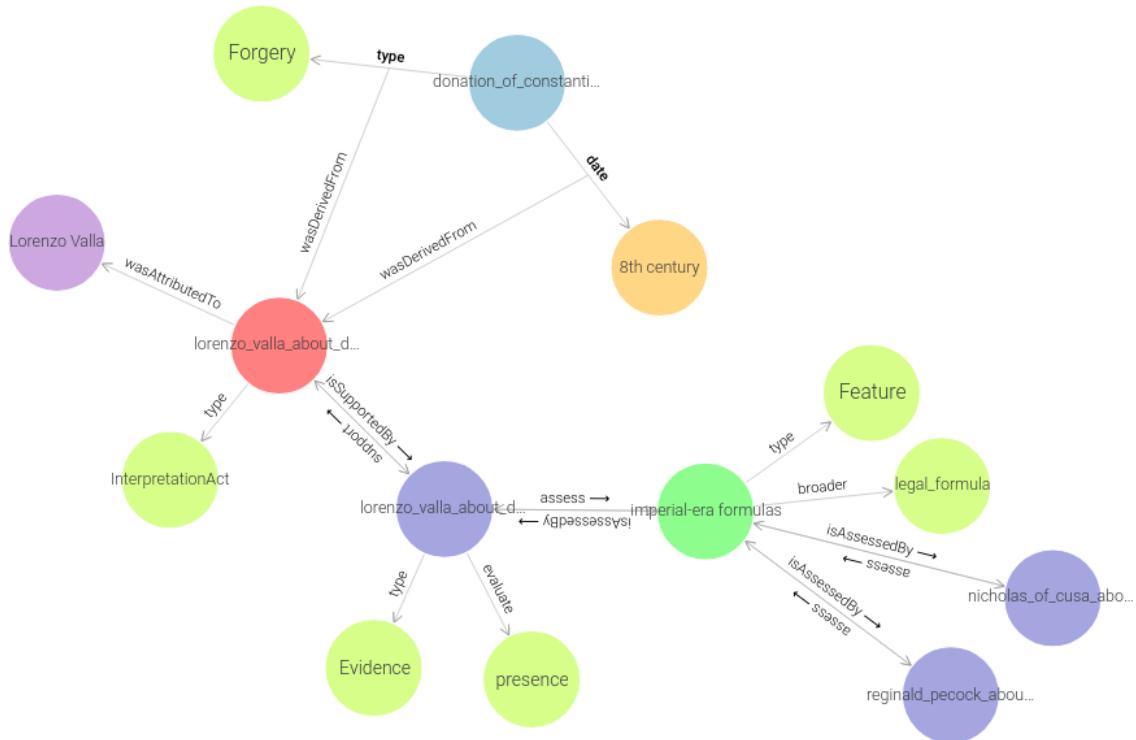


Figure 21: Lorenzo Valla’s statement about the Donation of Constantine

```

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.

Listing 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 .

```

### 5.3 Evaluation Framework

Our evaluation framework provides a multi-dimensional assessment of the KG generation pipeline, as described in Section 3.6, evaluating 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. The framework was implemented to systematically addresses five Evaluation Questions (EQs), in line with RQ2 to RQ5 (See 1).

**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 Sec 5.1.1 against the 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<sub>1</sub>-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 5.1.2.

**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 4.2), 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 requirements

**Scope:** Evidence evaluation is restricted to entities successfully matched between model output and GT from RQ2.

**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<sub>1</sub>-score evaluation framework established for RQ1 to assess the hypothesis extraction component described in Section 5.1.6. 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 RQ2 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 fidelity of the representation we decided to use G-EVAL (Liu et al., 2023). 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 in (Gangemi et al., 2024). We chose 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 relationships within the discourse narrative. It must be taken into account that previous evaluation mostly covered matchable entries between GT and output. G-EVAL evaluates the whole output.

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

## 6 Results

This section presents a comprehensive evaluation and preliminary discussion of findings across the five evaluation questions (EQs) outlined in Section 5.3. 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 22.

Table 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

Table 3 and Image 23 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

## 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 22: SPARQL query used to extract entity counts from the KGs for statistical comparison across models

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 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.

### 6.1 EQ1: CH Item Metadata Extraction Precision

**How accurately does the pipeline extract alleged CH Item metadata compared to expert annotations?**

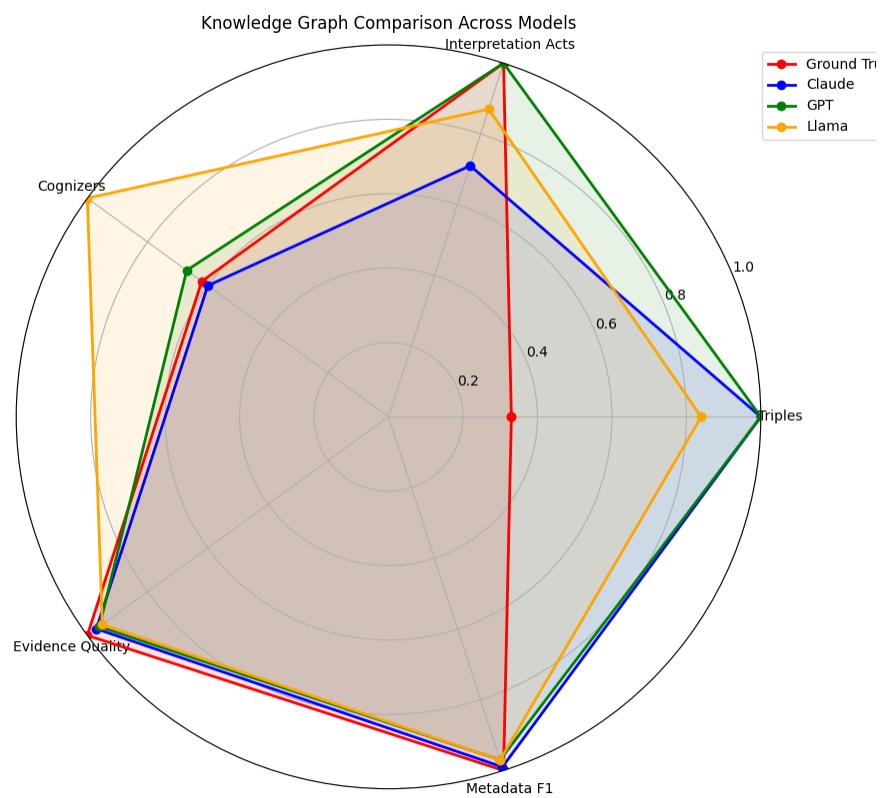
As shown in Table 4, the performance is high across all models, with F<sub>1</sub>-scores ranging from 0.97 to 0.99.

Claude Sonnet 3.7 achieves the highest overall performance with an F<sub>1</sub>-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.

### 6.2 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 3, the number of Cognizers is relatively similar across models - Table 5 shows the number of overlapping entities between the model's KG and the GT.

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%.



Listing 23: Radar Chart of different KG extractions

Table 4: CH Item Metadata extraction performance across three LLMs

<b>Model</b>	<b>Category</b>	<b>Precision</b>	<b>Recall</b>	<b>F<sub>1</sub>-Score</b>
<b>Claude</b>	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
	<b>Overall</b>	<b>0.987</b>	<b>0.995</b>	<b>0.991</b>
<b>GPT</b>	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
	<b>Overall</b>	<b>0.942</b>	<b>1.000</b>	<b>0.970</b>
<b>Llama</b>	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
	<b>Overall</b>	<b>0.933</b>	<b>1.000</b>	<b>0.965</b>

Table 5: Entity recognition coverage and accuracy

<b>Model</b>	<b>Precision</b>	<b>Recall</b>	<b>F<sub>1</sub></b>	<b>TP</b>	<b>FP</b>	<b>FN</b>
<b>Claude</b>	0.696	0.763	0.728	71	31	22
<b>GPT</b>	<b>0.718</b>	<b>0.912</b>	<b>0.803</b>	145	57	14
<b>Llama</b>	0.626	0.817	0.709	107	64	24

### 6.3 EQ3: Evidential Reasoning Extraction Quality

How accurately does the model capture the multi-dimensional evidential reasoning employed by scholars in their interpretations?

Table 6 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.

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 5, Claude shows the highest evidence coverage (0.968) in Table 6. This pattern highlights that the lower recall in identifying Cognizers by Claude returns in higher precision in downstream tasks.

### 6.4 EQ4: Hypothesis and Judgment Identification

How accurately does the model extract scholars' interpretative hypotheses and overall authenticity judgments?

Table 7 presents performance on extracting scholarly hypotheses about items origins and au-

Table 6: Evidence extraction quality and coverage

Model	Mean Score (0-4)	Percentage Score (%)
Claude	<b>3.87</b>	<b>96.8</b>
GPT-4o-mini	3.84	96.0
Llama 3.3	3.81	95.3

thenticity judgments.

Table 7: Hypothesis and judgment extraction performance

Model	Macro	Type	Creator	Date	Location
	F <sub>1</sub>				
Claude	0.655	0.652	0.638	<b>0.791</b>	<b>0.923</b>
GPT	<b>0.749</b>	<b>0.845</b>	0.484	0.595	0.727
Llama	0.694	0.691	<b>0.712</b>	0.762	0.727

GPT-4o-mini achieves the highest overall F<sub>1</sub>-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<sub>1</sub>) and temporal hypotheses (0.791 F<sub>1</sub>), 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<sub>1</sub>).

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.

## 6.5 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 evaluation 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 8: Per-statement Correctness (G-EVAL scores on 0-1 scale)

Model	Mean	Std Dev	Range
Claude	<b>0.620</b>	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 9: Overall Debate Representativeness (G-EVAL scores on 0-1 scale)

Model	Mean	Std Dev	Range
Claude	<b>0.607</b>	0.121	0.333 - 0.889
GPT	0.580	0.199	0.222 - 0.889
Llama	0.523	0.144	0.222 - 0.778

## 7 Discussion and Conclusions

In this section, we discuss the overall performance patterns, identified bottlenecks, and potential steps to enhance the KE while answering our RQs (Section 1), followed by our contributions, limitations and future steps.

### 7.1 Methodological Framework Validation

To answer RQ1, our five-step ATR4CH methodology proves effective in coordinating LLM-based extraction with ontological frameworks. The 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 FPs and FNs without reducing the necessity of external validation.

### 7.2 Extraction Performance Analysis

To answer RQ2, 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.

### 7.3 Representation Fidelity and Quality Assessment

To answer RQ3, 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 (Pasqual, 2025), 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.

### 7.4 Model Comparison and Performance Trade-offs

To answer RQ4, 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 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.

### 7.5 Deployment Implications and Cost-Effectiveness

To answer RQ5, the performance differences between models are relatively modest, while model sizes and costs differ substantially<sup>32</sup>. This suggests that the step-by-step pipeline architecture effectively leverages the capabilities of smaller 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

<sup>32</sup>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.

projects can leverage automated scaling through extensive annotation datasets and local deployment.

## 7.6 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 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 works that try to summarize

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.

## Acknowledgments

For camera-ready Acknowledge funding sources and contributors.

**AI Tool Disclosure:** This research employed Large Language Models (Claude Sonnet 3.7, Llama 3.3 70B, and GPT-4o-mini) as research subjects for knowledge extraction experiments, as detailed in the methodological sections of this article. Additionally, G-EVAL, an LLM-based evaluation framework, was used for assessing discourse representation quality. No AI tools were used in the writing, analysis, or interpretation of this manuscript beyond the specified cases above. The authors maintain full responsibility for the research design, methodology, data interpretation, figure creation, and all conclusions presented.

## References

- Bradley P. Allen, Lise Stork, and Paul Groth. Knowledge Engineering Using Large Language Models. *Transactions on Graph Data and Knowledge*, 1(1):3:1–3:19, 2023. ISSN 2942-7517. doi: 10.4230/TGDK.1.1.3. URL <https://drops.dagstuhl.de/entities/document/10.4230/TGDK.1.1.3>.

T. Andrews. The structured assertion record (star) model for event-based representation of historical information. In *GrapHNR 2023*, Mainz, Germany, 2023. URL <https://graphentechnologie.hypotheses.org/files/2023/05/GrapHNR-2023-32-Andrews-STAR.pdf>.

Luigi Asprino, Enrico Daga, Aldo Gangemi, and Paul Mulholland. Knowledge graph construction with a façade: A unified method to access heterogeneous data sources on the web. *ACM Trans. Internet Technol.*, 23(1), 2023. ISSN 1533-5399. doi: 10.1145/3555312. URL <https://doi.org/10.1145/3555312>.

Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

G. Barabucci, F. Tomasi, and F. Vitali. Supporting complexity and conjectures in cultural heritage descriptions. In *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, pages 104–115. CEUR Workshop, 2021.

S. Baroncini, B. Sartini, M. van Erp, F. Tomasi, and A. Gangemi. Is dc:subject enough? a landscape on iconography and iconology statements of knowledge graphs in the semantic web. *Journal of Documentation*, 79:115–136, 2023. doi: 10.1108/JD-09-2022-0207.

N. Barone. Intorno alla falsificazione dei documenti ed alla critica di essi. memoria letta all'accademia pontaniana nella tornata del 21 gennaio 1912. *Atti Dell'Accademia Pontaniana*, 42, 1912. URL <http://www.rmoa.unina.it/4359/>.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. MEDEC: A benchmark for medical error detection and correction in clinical notes. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22539–22550, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1159. URL <https://aclanthology.org/2025.findings-acl.1159/>.

E. Bernasconi and S. Ferilli. New frontiers in digital libraries: The trajectory of digital humanities through a computational lens. In *3rd Workshop on Artificial Intelligence for Cultural Heritage (AI4CH 2024)*, AI4CH 2024, Bolzano, Italy, 2024. doi: 10.5281/zenodo.14923857. URL <https://ai4ch.di.unito.it/>.

N. Blau. Uncertainty and the history of ideas. *History and Theory*, 50(3):358–372, 2011. doi: 10.1111/j.1468-2303.2011.00590.x.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020a. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020b. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.

Valentina Anita Carriero, Fabio Mariani, Andrea Giovanni Nuzzolese, Valentina Pasqual, and Valentina Presutti. Agile knowledge graph testing with testalod. In *ISWC (Satellites)*, pages 221–224, 2019. URL <https://ceur-ws.org/Vol-2456/paper58.pdf>.

J.J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs. *Journal of Web Semantics*, 3(4):247–267, 2005. doi: 10.1016/j.websem.2005.09.001.

Peter Checkland and Sue Holwell. Data, capta, information and knowledge. In *Introducing Information Management: the business approach*, pages 47–55. Elsevier, 2006. ISBN 0-7506-6668-4. doi: 10.4324/9780080458397-10.

Marilena Daquino and Francesca Tomasi. Historical context ontology (hico): A conceptual model for describing context information of cultural heritage objects. In E. Garoufallou, R. Hartley, and P. Gaitanou, editors, *Metadata and Semantics Research. MTSR 2015*, volume 544 of *Communications in Computer and Information Science*, Cham, 2015. Springer. doi: 10.1007/978-3-319-24129-6\_37.

Marilena Daquino, Valentina Pasqual, and Francesca Tomasi. Knowledge representation of digital hermeneutics of archival and literary sources. *JLIS.It*, 11(3):59–76, 2020. doi: 10.4403/jlis.it-12642.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423/>.

Alessio Di Pasquale, Valentina Pasqual, Francesca Tomasi, and Fabio Vitali. On assessing weaker logical status claims in wikidata cultural heritage records. *Semantic Web Journal*, 2024.

Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. RML: a generic language for integrated RDF mappings of heterogeneous data. In Christian Bizer, Tom Heath, Sören Auer, and Tim Berners-Lee, editors, *Proceedings of the 7th Workshop on Linked Data on the Web*, volume 1184 of *CEUR Workshop Proceedings*, 2014. URL <http://ceur-ws.org/Vol-1184/lidow2014%5Fpaper%5F01.pdf>.

Martin Doerr. The cidoc conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75, 2003. doi: 10.1609/aimag.v24i3.1720. URL <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1720>.

- Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 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)*, pages 49–54, Baltimore, Maryland, 2014. Association for Computational Linguistics. URL <https://aclanthology.org/P14-2009>.
- Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Hassan El-Hajj and Matteo Valleriani. Cidoc2vec: Extracting information from atomized cidoc-crm humanities knowledge graphs. *Information*, 12(12), 2021. ISSN 2078-2489. doi: 10.3390/info12120503. URL <https://www.mdpi.com/2078-2489/12/12/503>.
- Hans G. Gadamer. *Truth and Method*. A&C Black, London, 2013.
- Aldo Gangemi, Arianna Graciotti, Eleonora Marzi, Antonello Meloni, Andrea Nuzzolese, Valentina Presutti, Diego Reforgiato Recupero, Alessandro Russo, and Rocco Tripodi. MusicBO, an application of Text2AMR2FRED to the Musical Heritage domain. In *20th Extended Semantic Web Conference*, Crete, Greece, 2024. CEUR Workshop Proceedings.
- Z. Gao, A. Feng, X. Song, and X. Wu. Target-dependent sentiment classification with bert. *IEEE Access*, 7:154290–154299, 2019. doi: 10.1109/ACCESS.2019.2946594.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1017. URL <https://www.aclweb.org/anthology/P17-1017.pdf>.
- Lucia Giagnolini, Andrea Schimmenti, Paolo Bonora, and Francesca Tomasi. Expliciting contexts: Semantic knowledge extraction from traditional archival descriptions. *Umanistica Digitale*, 9(20):115–144, 2025. doi: 10.6092/issn.2532-8816/21229. URL <https://umanisticadigitale.unibo.it/article/view/21229>.
- S. Haider. Verzeichnis der den oberösterreichischen raum betreffenden gefälschten, manipulierten oder verdächtigten mittelalterlichen urkunden. Technical report, Oberösterreichisches Landesarchiv, 2022.
- F. Hamborg, K. Donnay, and B. Gipp. Towards target-dependent sentiment classification in news articles. In K. Toeppe, H. Yan, and S.K.W. Chu, editors, *Diversity, Divergence, Dialogue. iConference 2021*, volume 12646 of *Lecture Notes in Computer Science*, pages 157–169, Cham, 2021. Springer. doi: 10.1007/978-3-030-71305-8\_12.
- R. Härtel. Il falso documento del conte giovanni di moggio (875). In G. Pugnetti and B. Lucci, editors, *Mueç. Societât Filologjiche Furlane/Società Filologica Friulana, XCIV Congrès*, pages 247–252, Udin/Udine, 2017.
- Olaf Hartig. Foundations of rdf\*and sparql\*:(an alternative approach to statement-level metadata in rdf). In Juan L. Reutter and Divesh Srivastava, editors, *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web*, volume 1912 of *CEUR Workshop Proceedings*. Juan Reutter, Divesh Srivastava, CEUR-WS.org, 2017. URL <http://ceur-ws.org/Vol-1912/paper12.pdf>.

Jie He, Yijun Yang, Wanqiu Long, Deyi Xiong, Victor Gutierrez Basulto, and Jeff Z. Pan. Evaluating and improving graph to text generation with large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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)*, pages 10219–10244, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.nacl-long.513. URL <https://aclanthology.org/2025.nacl-long.513/>.

IFLA Working Group on FRBR/CRM Dialogue. Definition of frbroo: A conceptual model for bibliographic information in object-oriented formalism. Technical report, International Federation of Library Associations and Institutions (IFLA), 2017. URL <https://repository.ifla.org/handle/20.500.14598/659>.

Hanieh Khorashadizadeh, Fatima Zahra Amara, Morteza Ezzabady, Frédéric Ieng, Sanju Tiwari, Nandana Mihindukulasooriya, Jinghua Groppe, Soror Sahri, Farah Benamara, and Sven Groppe. Research Trends for the Interplay between Large Language Models and Knowledge Graphs, 2024. URL <http://arxiv.org/abs/2406.08223>.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The INCEption platform: Machine-assisted and knowledge-oriented interactive annotation. In Dongyan Zhao, editor, *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico, 2018. Association for Computational Linguistics. URL <https://aclanthology.org/C18-2002/>.

Vivek Kumar, Diego Reforgiato Recupero, Rim Helaoui, and Daniele Riboni. K-LM: Knowledge Augmenting in Language Models Within the Scholarly Domain. *IEEE Access*, 10:91802–91815, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3201542.

Yassir Lairgi, Ludovic Moncla, Rémy Cazabet, Khalid Benabdeslem, and Pierre Cléau. iText2KG: Incremental Knowledge Graphs Construction Using Large Language Models, 2024. URL <http://arxiv.org/abs/2409.03284>.

T. Lebo et al. Prov-o: The prov ontology. W3c recommendation, World Wide Web Consortium, 2013. URL <http://www.w3.org/TR/2013/REC-prov-o-20130430/>.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS ’20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Pasquale Lisena, Daniel Schwabe, Marieke van Erp, Raphaël Troncy, William Tullett, Inger Lee-mans, Lizzie Marx, and Sofia Colette Ehrich. Capturing the Semantics of Smell: The Odeuropa Data Model for Olfactory Heritage Information. In Paul Groth, Maria-Esther Vidal, Fabian Suchanek, Pedro Szekley, Pavan Kapanipathi, Catia Pesquita, Hala Skaf-Molli, and Minna Tamper, editors, *The Semantic Web*, pages 387–405, Cham, 2022. Springer International Publishing. ISBN 978-3-031-06981-9.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

*Processing*, pages 2511–2522, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL <https://aclanthology.org/2023.emnlp-main.153/>.

Diana Maynard, Kalina Bontcheva, and Isabelle Augenstein. *Natural Language Processing for the Semantic Web*. Synthesis Lectures on Data, Semantics, and Knowledge. Springer Cham, Cham, Switzerland, 2017. ISBN 978-3-031-79473-5. doi: 10.1007/978-3-031-79474-2. URL <https://doi.org/10.1007/978-3-031-79474-2>.

Antonello Meloni, Diego Reforgiato Recupero, and Aldo Gangemi. Amr2fred, a tool for translating abstract meaning representation to motif-based linguistic knowledge graphs. In *Extended Semantic Web Conference*, 2017. URL <https://api.semanticscholar.org/CorpusID:34725770>.

L.-P. Meyer, C. Stadler, et al. Llm-assisted knowledge graph engineering: Experiments with chatgpt. In C. Zinke-Wehlmann and J. Friedrich, editors, *First Working Conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow. AIDRST 2023*, Informatik aktuell, pages 157–169, Wiesbaden, 2024. Springer Vieweg. doi: 10.1007/978-3-658-43705-3\_8.

N. Mihindukulasooriya, S. Tiwari, E.C. Fernández, and K. Lata. Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. In Terry R. Payne, Duy Dinh Huynh, Joongmin Kim, Hala Haddad, Zareen Afzal, Jeff Z. Pan, Mark Chapman, Fabien L. Gandon, Rohit Krishna, Michel Dumontier, and Jun Zhao, editors, *The Semantic Web – ISWC 2023*, volume 14266 of *Lecture Notes in Computer Science*, Cham, 2023. Springer. doi: 10.1007/978-3-031-47243-5\_14.

S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.emnlp-main.759>.

Alba Morales Tirado, Jason Carvalho, Paul Mulholland, and Enrico Daga. Musical Meetups: a Knowledge Graph approach for Historical Social Network Analysis. In Mehwish Alam, Cassia Trojahn, Sven Hertling, Catia Pesquita, Christian Aebeloe, Hidir Aras, Amr Azzam, Juan Cano, John Domingue, Simon Gottschalk, Olaf Hartig, Katja Hose, Sabrina Kirrane, Pasquale Lisena, Francesco Osborne, Philipp Rohde, Luc Steels, Ruben Taelman, Aisling Third, Ilaria Tiddi, and Rima Türker, editors, *ESWC 2023 Workshops and Tutorials. Semantic Methods for Events and Stories (SEMMES)*, volume 3443. CEUR Workshop Proceedings (CEUR-WS.org), 2023. URL <https://ceur-ws.org/Vol-3443/ESWC%5f2023%5fSEMMES%5fMeetups-CR.pdf>.

Sarah Rebecca Ondraszek, Grischka Petri, Ulrike Blumenthal, Lisa Dieckmann, Etienne Posthumus, and Harald Sack. eXtreme Design for Ontological Engineering in the Digital Humanities with Viewsari, a Knowledge Graph of Giorgio Vasari’s The Lives. In *Proceedings of the 4th Workshop on Semantic Web Technologies for Digital Humanities (SemDH 2024)*, 2024. URL <https://ceur-ws.org/Vol-3724/paper5.pdf>.

Ontotext. Graphdb: Semantic database, 2024. URL <https://www.ontotext.com/products/graphdb/>.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.

Valentina Pasqual. *The Critical Inquiry in Humanities Knowledge Graphs: Challenges, Methods, Innovations*. PhD thesis, alma, 2025.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL <https://aclanthology.org/D19-1250/>.

M. Piotrowski. Uncertainty as unavoidable good. *Universität Bielefeld, Center for Uncertainty Studies (CeUS)*, 5:10, 2023. doi: 10.4119/unibi/2983506. URL <https://pub.uni-bielefeld.de/record/2983506>.

M. Piotrowski and M. Neuwirth. Prospects for computational hermeneutics. In *Proceedings of the 9th AIUCD Annual Conference*, 2020. URL <http://amsacta.unibo.it/6316/>.

M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, 2014. Association for Computational Linguistics. URL <https://aclanthology.org/S14-2004>.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, 2015. Association for Computational Linguistics.

Valentina Presutti, Enrico Daga, Aldo Gangemi, and Eva Blomqvist. extreme design with content ontology design patterns. In *Proceedings of the 2009 International Conference on Ontology Patterns - Volume 516*, WOP’09, page 83–97, Aachen, DEU, 2009. CEUR-WS.org.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Guoliang Li, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models. *ACM Comput. Surv.*, 57(4), 2024. ISSN 0360-0300. doi: 10.1145/3704435. URL <https://doi.org/10.1145/3704435>.

Célian Ringwald. Learning Pattern-Based Extractors from Natural Language and Knowledge Graphs: Applying Large Language Models to Wikipedia and Linked Open Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23411–23412, 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i21.30406. URL <https://ojs.aaai.org/index.php/AAAI/article/view/30406>.

B. Sartini, S. Baroncini, M. van Erp, F. Tomasi, and A. Gangemi. Icon: An ontology for comprehensive artistic interpretations. *J. Comput. Cult. Herit.*, 16(3):59–76, 2023. doi: 10.1145/3594724.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=Yacmpz84TH>.

Gytė Tamasauskaitė and Paul Groth. Defining a knowledge graph development process through a systematic review. *ACM Transactions on Software Engineering and Methodology*, 32:1–40, 2022. URL <https://api.semanticscholar.org/CorpusID:248435579>.

Francesca Tomasi. Digital humanities e organizzazione della conoscenza: una pratica di insegnamento nel lodlam. *AIB STUDI*, 60(2):411–425, 2020. doi: 10.2426/aibstudi-12068. URL <https://aibstudi.aib.it/article/view/12068>.

L. Valla. *The treatise of Lorenzo Valla on the Donation of Constantine*. New Haven: Yale University Press, 2023. URL <https://www.gutenberg.org/ebooks/70092>.

Z. Wang, Q. Xie, Z. Ding, Y. Feng, and R. Xia. Is chatgpt a good sentiment analyzer? a preliminary study, 2023. URL <https://api.semanticscholar.org/CorpusID:258048703>.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. BartScore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nacl-long.300. URL <https://aclanthology.org/2024.nacl-long.300>.