Received 27 July 2025
Revised 29 January 2026
Accepted 2 February 2026

Knowledge graphs generation from cultural heritage texts: combining LLMs and ontological engineering for scholarly debates

Andrea Schimmenti
Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
Valentina Pasqual
Department of Classical Philology and Italian Studies, Digital Humanities Advanced Research Center (/DH.arc), University of Bologna, Bologna, Italy
Fabio Vitali
Department of Computer Science and Engineering, University of Bologna, Bologna, Italy, and
Marieke van Erp
DHLab, KNAW Humanities Cluster, Amsterdam, Netherlands

Abstract

Purpose – Cultural heritage (CH) 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 (LLM)-based knowledge extraction from CH documents. We validate the methodology through a case study on authenticity assessment debates.

Design/methodology/approach – 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, etc.), implementing a sequential pipeline with three LLMs (Claude Sonnet 3.7, Llama 3.3 70B and GPT-4o-mini).

Findings – The methodology successfully extracts complex CH knowledge: 0.96–0.99 F1 for metadata extraction, 0.7–0.8 F1 for entity recognition, 0.65–0.75 F1 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.

Research limitations/implications – The produced KG is limited to Wikipedia articles. While the results are encouraging, human oversight is necessary during post-processing.

Originality/value – To the best of the authors' knowledge, this is the first systematic methodology for coordinating LLM-based extraction with CH ontologies. ATR4CH provides a replicable framework adaptable across CH domains and institutional resources. ATR4CH enables CH 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 (LLMs)

Paper type Research article

Journal of Documentation
Vol. 82 No. 7, 2026
pp. 206-250
Emerald Publishing Limited
e-ISSN: 1758-7379
p-ISSN: 0022-0418
DOI 10.1108/JD-07-2025-0203

Funding: Research partially funded by the European Union -Next Generation EU, investment I.4.1 PNRR Patrimonio Culturale, Decreto Ministeriale n. 351 del 9 aprile 2022.

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 concentrated for the most part 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 knowledge contained in unstructured texts (descriptive content, contextual information and analytical discourse) remains difficult to extract and structure systematically into queryable formats, and even when integrated into KGs, it is usually kept in 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. Additional challenges 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 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 are treated as an objective and context-independent representation of reality (Peter and Holwell, 2006). These epistemological tensions manifest across various forms of CH scholarship, from attribution studies and provenance research to interpretative analysis and critical evaluation. In authenticity assessment, scholars from different humanities disciplines (e.g., Diplomatics, Paleography, 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). However, current KG implementations represent only simplified versions of scholarly discourse. 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 [1] and DBpedia [2] 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. Consider the famous *Donation of Constantine*, a 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 the 4th century. Despite Valla's compelling evidence, acceptance of this finding evolved gradually over centuries.

As shown in Figure 1, Wikidata categorizes the *Donation* as a "historical forgery" [3] with no representation of the scholarly debate, while DBpedia [4] 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. Three interconnected challenges exist. The first is *syntactic*: representing competing scholarly opinions within formal knowledge representation (KR) 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

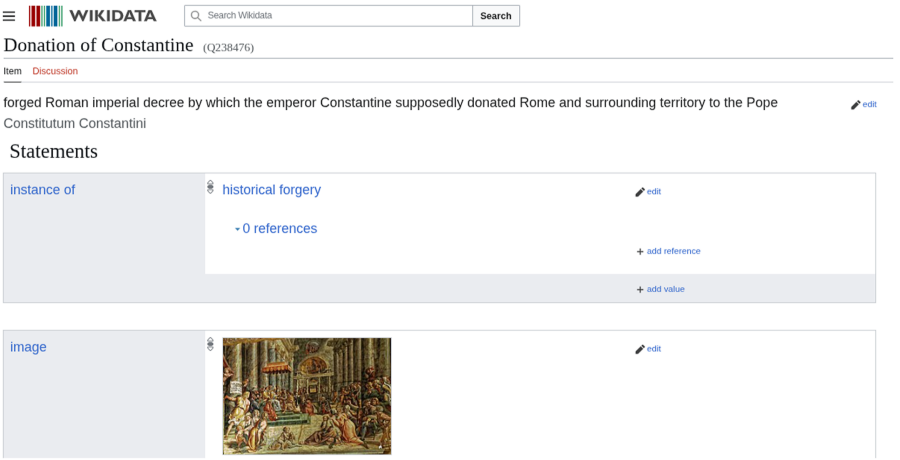

Figure 1. The *Donation of Constantine* entry in Wikidata. Source: Wikidata

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 expert annotators to identify scholarly agents, extract evidential reasoning and capture alternative hypotheses while maintaining consistency across large document collections. 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 ability to process complex academic discourse, identify implicit relationships and handle domain-specific vocabularies (Khorashadizadeh *et al.*, 2024) without additional training. LLMs can bootstrap knowledge extraction (KE) pipelines by eliminating the need for specific annotated training corpora and exploiting transfer learning across domains beyond those on which they were explicitly trained (Brown *et al.*, 2020).

The third challenge is *methodological*: despite advances in both ontology development and extraction techniques, the literature lacks systematic methodologies that integrate ontology-driven KG generation specifically for CH contexts. While general-purpose text-to-KG frameworks exist (Maynard *et al.*, 2017; Hotho *et al.*, 2020), they typically assume the availability of large annotated training corpora and domain-agnostic entity types. What remains absent is a comprehensive methodological framework that guides practitioners through the complete process: from analyzing source materials and identifying extractable patterns, through developing appropriate annotation schemas and extraction pipelines, to evaluation. Such a methodology must accommodate the resource constraints typical of CH institutions while maintaining the representational fidelity required for humanities scholarship.

RQ. This work tackles the following primary research question: How can a systematic methodology integrate LLM-based 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 integrate 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 KGs 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?

Our contributions are threefold. First, we present the ATR4CH (Adaptive Text-to-RDF for Cultural Heritage) methodology, which combines annotation development, ontological alignment and pipeline-based extraction. The methodology is a replicable framework that can be adapted across CH domains and institutional resources. Second, we demonstrate the practical implementation by connecting ATR4CH to the Scholarly Evidence Based Interpretation (SEBI) ontology [5] (Pasqual, 2025) to develop an annotation model and an extraction pipeline 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.

The remainder of the paper is organized as follows: Section 2 reviews related works in KR, extraction methods for the Semantic Web, opinion mining and LLM-based KE approaches. Section 3 presents the ATR4CH methodology, detailing the five-task iterative approach for integrating LLM-based extraction with ontological frameworks in the CH domain. Section 4 describes ATR4CH implementation for the authenticity assessment use case, following the five-task structure: Section 4.1 presents foundational analysis and design, including corpus collection, the SEBI ontology and preliminary analyses; Section 4.2 describes iterative annotation schema development and ground truth (GT) preparation; Section 4.3 details pipeline architecture development; Section 4.4 and Section 4.5 describe pipeline refinement, KG generation and comprehensive evaluation. Section 5 provides experimental results across five evaluation questions (EQs), comparing 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. Section 6 discusses findings in relation to research questions, analyzes performance trade-offs, addresses deployment implications and outlines contributions, limitations and future directions. All code is available at https://github.com/aschimmenti/SEBI-Knowledge-Extraction.

2. Related work

The challenges of representing and extracting interpretative knowledge in the CH domain have received increasing attention in recent research. 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.

2.1 KR of certainty in the Semantic Web

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 digitization 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. 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) [6] 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. Other knowledge bases like YAGO4 [7] integrate RDF-Star to model provenance (e.g., for temporal information) but lack statements on the certainty of the given triple (Govindapillai *et al*., 2021). Despite the adequate expressive power made available by the Wikidata model, annotators in the CH domain underutilize 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 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. Previous work has introduced the SEBI ontology [8] Pasqual (2025), aimed at representing scholarly claims on a hand-curated catalogue of forged manuscripts, which resulted in the BROAST application [9].

2.2 KE for the Semantic Web

KE for the Semantic Web transforms unstructured textual content into machine-processable representations conforming to established ontological models. The text-to-KG task refers to systems that process natural language text to generate RDF KGs, where extraction is guided by predefined ontologies and outputs conform to ontological constraints (Maynard *et al*., 2017). This approach, termed Closed KE or Ontology-Based Information Extraction (Wimalasuriya and Dou, 2010), operates within predefined ontological frameworks specifying permitted entity types, relations and semantic constraints.

The Closed KE paradigm aligns with the Semantic Web vision, as extraction targets conform to ontological schemas designed for machine reasoning and cross-system integration (Berners-Lee *et al*., 2001). The requirements for ontology-oriented extraction systems manifest across three critical dimensions (Hotho *et al*., 2020): entity mentions must be mapped to URIs serving as globally unique identifiers within the LOD cloud, elevating Entity Linking from optional refinement to fundamental requirement; data alignment challenges emerge across T-Box alignment (schema extension for novel entity types), A-Box alignment (deduplication and consistency checking) and URI alignment (entity resolution against existing knowledge base entries); provenance tracking and validation mechanisms must ensure generated RDF conforms to both syntactic requirements and semantic constraints defined in the ontology. Text-to-KG systems following the Closed KE paradigm typically decompose extraction into specialized subtasks executed sequentially:

preprocessing and sentence segmentation, Named Entity Recognition (NER) to identify and classify entity mentions, Entity Linking to map mentions to canonical identifiers, Relation Extraction to identify semantic relationships, Event Detection when using event-centric ontologies and graph assembly with validation (Maynard *et al*., 2017). Recent CH projects demonstrate diverse pipeline implementation approaches. The Musical Meetups Knowledge Graph (Alba Morales *et al*., 2023) combines DBpedia Spotlight entity recognition with LLMs (GPT-3.5-Turbo) for temporal normalization and purpose classification. MusicBO (Gangemi *et al*., 2024) employs Abstract Meaning Representation as an intermediate layer (Meloni *et al*., 2017) between linguistic structure and ontological requirements. The Odeuropa project (Lisena *et al*., 2022) achieves tight annotation–ontology integration through frame-based schemes mapping directly to CIDOC CRM extensions using multilingual BERT models.

LLMs have introduced new text-to-KG paradigms (Khorashadizadeh *et al*., 2024; Mihindukulasooriya *et al*., 2023; Meyer and Stadler, 2024). Allen *et al*. (2023) identify hybrid neuro-symbolic systems and natural language interfaces as primary architectural directions, the latter enabling domain experts to guide extraction without technical expertise in KR. Lairgi *et al*.'s iText2KG (Lairgi *et al*., 2024) employs a zero-shot, incremental approach enabling knowledge base expansion without annotated training data. Ringwald (2024) explores pattern-based methods for learning from Wikipedia-DBpedia/Wikidata pairs. For CH-specific applications, Santini (2024) shows LLM-based relation extraction outperforming specialized models like mREBEL on 19th-century Italian texts through broader linguistic knowledge. Schimmenti *et al*. (2024) demonstrate zero-shot tools like GLiNER (Zaratiana *et al*., 2024), enabling on-the-fly entity type specification through natural language descriptions. Giagnolini *et al*. (2025) employ Llama 3.3 70B for text-to-KG extraction from archival metadata, classifying paragraphs by event types before applying event-specific extraction schemas mapped to RiC-O [10].

The integration of LLMs alters resource requirements and performance characteristics. LLMs leverage In-Context Learning (ICL) (Brown *et al*., 2020), Few-Shot learning strategies (Petroni *et al*., 2019) and Chain-of-Thought (CoT) prompting to perform extraction with minimal task-specific examples. This capability proves valuable for domains where creating large annotated corpora is often infeasible. On the other hand, LLM-based approaches introduce new challenges: ontological compliance requires sophisticated prompt engineering or post-processing, hallucination risks necessitate careful verification mechanisms and the implicit knowledge encoded in model parameters may not align with domain-specific conceptual frameworks embodied in CH ontologies.

Evaluation methodologies for generated KGs remain heterogeneous and often inadequate for capturing semantic fidelity beyond surface-level metrics. Traditional precision–recall calculations on entity mentions and relations fail to assess whether extracted knowledge structures adequately represent the complexity and nuance of scholarly discourse. Back-translation approaches (Gangemi *et al*., 2024) and LLM-based evaluation frameworks like G-EVAL (Liu *et al*., 2023) offer promising directions, but their applicability across different types of CH interpretative content requires further validation.

3. The ATR4CH methodology

This section introduces the *ATR4CH* methodology, an iterative approach for extracting KGs from CH documents using LLMs. ATR4CH recognizes the fundamental interdependence of annotation, KE and ontology alignment, approaching them conjunctively.

3.1 Methodology overview

The ATR4CH methodology transforms three foundational inputs – unstructured document corpus, target ontology and Competency Questions (CQs) – into a validated KE system

through five interconnected tasks, producing a working extraction pipeline, refined annotation model and comprehensive evaluation framework.

A flowchart is shown in Figure 2. The methodology draws on eXtreme Design (Presutti *et al.*, 2009) for iterative ontology engineering centered on CQs, selection methodologies

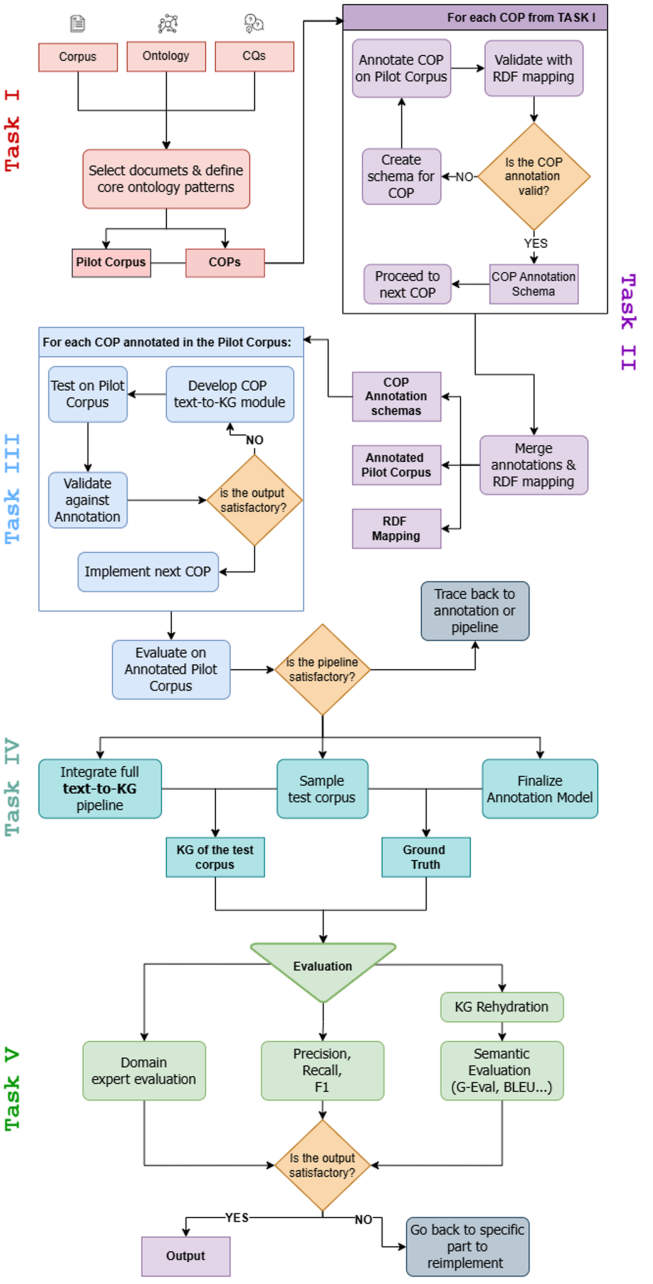

Figure 2. Flowchart of the ATR4CH methodology showing the iterative task structure. Source: Authors' own work

(Tomasi, 2020) and Odeuropa's integrated annotation–ontology approach (Lisena *et al*., 2022).

ATR4CH focuses on integrating annotation and KE pipeline development with ontologies in the CH domain, including CIDOC CRM (Doerr, 2003), Dublin Core [11], FRBR/FRBRoo (IFLA Working Group on FRBR/CRM Dialogue, 2017), HiCO (Daquino and Tomasi, 2015), SKOS [12] and PROV-O (Lebo *et al*., 2013). The methodology presupposes dual alignment: the ontology must represent relevant domain knowledge, and this knowledge must be present (explicitly or inferably) within source documents. It suits unstructured texts (informative, narrative, scholarly sources) rather than semi-structured documents like catalogues. The methodology leverages LLM capabilities in ICL (Brown *et al*., 2020), Few-Shot and CoT strategies (Petroni *et al*., 2019; Lairgi *et al*., 2024), based on established KE practices (Tamasauskaitė and Groth, 2022).

ATR4CH adopts an *incremental, pattern-by-pattern development strategy,* iteratively focusing on one Core Ontological Pattern (COP) at a time, identified through target ontology and CQ analysis. Each COP progresses through the complete development cycle – annotation schema design, RDF mapping validation and automated extraction module implementation – before advancing to the next pattern. After several pattern-specific iterations, the methodology transitions from the *Pilot Corpus phase* to full-scale corpus processing. The annotation model is consolidated into a production-ready version to work as the GT for pipeline evaluation.

3.2 Foundational analysis and design (Task I)

Task I establishes foundational understanding by analyzing the corpus and ontology to identify COPs, addressing data sparseness common in Information Extraction from unstructured texts.

(1) Corpus Analysis: This ontology-dependent activity examines knowledge manifestation in textual discourse, including linguistic patterns, discourse structures and representational strategies. Challenges include implicit mentions requiring contextual inference, long-distance dependencies where KG components are separated by substantial text spans, nested entities in relational structures and ambiguous references. For Wikipedia articles about forged CH items, structural analysis identifies which sections contain scholarly opinions versus tangential debates, enabling focused extraction from high-density sections like "Scholarly analysis." Content analysis determines whether articles present complete scholarly reasoning or merely final judgments, guiding the methodology toward sources with sufficient depth.

(2) Ontology Analysis: This parallel activity assesses which ontology parts can be populated from source documents, examining alignment between the ontology's conceptual framework and available textual information. It identifies which classes and properties have sufficient textual evidence, which relationships can be inferred from corpus patterns and which elements need omission. The aim is to determine *what* data are present rather than immediately addressing *how* to extract. CQs guide prioritization of ontological coverage based on research requirements.

(3) COPs Identification: Based on analyses, this process identifies essential KG patterns required to answer CQs. COPs represent central ontological nodes and relationships that are both present as extractable information and necessary for addressing research questions. Patterns emerge through systematic intersection analysis of CQs, ontological capabilities and textual evidence. Each candidate pattern is evaluated on *necessity* (required for answering CQs?) and *feasibility* (sufficient textual evidence for reliable extraction?). Patterns scoring high on both dimensions form the initial set, refined by considering dependencies and complexity. Simpler patterns are prioritized for early iterations to establish baseline functionality. The final selection represents a manageable subset forming the semantic backbone for KE, with patterns ordered by

their structural role – foundational metadata patterns preceding interpretative reasoning patterns. These COPs will be processed incrementally through subsequent tasks.

(4) Pilot Corpus Selection: The Pilot Corpus is a representative document set serving as a development sandbox. It is not a quantitatively representative sample but a qualitative one that must be linguistically, structurally and epistemically representative while remaining manageable for intensive manual work. Selection ensures coverage of linguistic patterns, discourse structures and diverse COP manifestations. Size can be three to five documents, depending on length and information complexity. This set will be used iteratively for developing and validating extraction pipelines for each COP.

3.3 Minimal working annotation development (Task II)

Task II develops annotation schemas incrementally, processing one COP at a time. For each COP identified in Task 3.2, an annotation schema is developed, applied to the Pilot Corpus and validated through RDF mapping before proceeding to the next pattern, producing an annotation model serving as the target schema for automated extraction.

Pattern-by-Pattern Annotation Schema Development: Development proceeds iteratively through identified COPs. For each pattern, an annotation schema captures essential knowledge structures while remaining practical for manual annotation and automated extraction. Schema design accounts for diverse knowledge manifestations in the corpus, including explicit textual mentions and information requiring inference or contextualization.

The annotation schema should prioritize simplicity and feasibility while ensuring adequate coverage. "Minimal" refers to including only necessary annotation elements for extracting identified ontological patterns in the first iterations, avoiding over-annotation that complicates extraction without contributing to answering CQs. If COPs require complex semantic structures beyond simple triple patterns, the annotation schema should include appropriate mechanisms for representing these relationships mappable to RDF (e.g., Named Graphs, reification).

Knowledge Base Integration Strategy: Knowledge base integration enables consistent entity identification and vocabulary alignment between textual mentions and the target ontology. Since COPs typically involve ontological individuals, entities, controlled vocabularies or standardized terminologies, annotators need access to these resources to ensure textual references link to correct ontological entities, preventing inconsistent annotation that would hamper aggregation and reasoning in the final KG.

Integration – whether through local vocabularies or external resources like Wikidata or DBpedia – must be designed early to establish clear protocols for entity linking and vocabulary alignment, guiding both manual annotation and automated extraction in Task 3.4. Choice between local and external knowledge bases depends on domain coverage, data quality requirements and specific entity types required by COPs.

Annotation Paradigm: Annotation should follow established corpus linguistics and NLP practices. When resources permit, multiple annotators should annotate the same documents to enable inter-annotator agreement measurement using Cohen's kappa (Carletta, 1996; Cohen, 1960) or Krippendorff's alpha (Krippendorff, 2019), identifying ambiguous categories and revealing where guidelines require clarification. In resource-constrained settings, a single experienced annotator may suffice, but annotation guidelines must be thoroughly documented for reproducibility. The process should be iterative: initial guidelines are refined based on encountered edge cases.

Iterative Development Process for Each COP: Development follows a systematic cycle, ensuring the annotation schema produces RDF structures satisfying COPs:

(1) Schema Design: Develop initial annotation layers based on the current COP, incorporating knowledge base integration protocols through tagsets, controlled vocabularies and standardized terminologies aligning with the target ontology.

(2) Pilot Corpus Annotation: Annotate the entire pilot corpus using the current schema iteration to identify gaps, inconsistencies or bottlenecks.

(3) Mapping Validation: Conduct preliminary mapping exercises from annotated data to RDF format, testing whether resulting KGs satisfy the COP and adequately represent source document semantic content. This mapping serves as a unit test, validating that annotation patterns correctly transform to valid RDF.

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

(5) Pipeline Development for Current COP: Once the annotation schema has been validated through successful RDF mapping, proceed to Task 3.4 to develop automated extraction for this pattern using the same Pilot Corpus documents. Only after completing pipeline development and validation for the current COP should development proceed to the next pattern.

These preliminary mapping exercises validate that the annotation schema produces target knowledge structures, serving as an early validation mechanism before proceeding to automated extraction development. The Minimal Working Annotation emerges as the aggregation of annotation schemas developed for individual COPs.

3.4 Pipeline architecture development (Task III)

Task III designs and implements computational tools to automatically extract COPs from text using the annotation model as the target schema, addressing the CH corpora's domain-specific characteristics and limited annotated training data. This task develops and validates extraction capabilities incrementally for each COP using the same Pilot Corpus documents annotated in the previous task.

Task Decomposition and Architecture Design: KE is designed around annotation model 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. For the current COP, the extraction pipeline targets the relative annotation schema developed in Task 3.3.

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

(1) Low data, low resources: API-based LLMs with few-shot prompting and rule-based entity linking

(2) Moderate data, moderate resources: Hybrid approaches combining pre-trained models with domain-specific fine-tuning

(3) Large data, extensive resources: Custom model training and ensemble methods

(4) 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.*, 2020; Min *et al.*, 2022), combined with specialized NER tools (Devlin *et al.*, 2019) for precise span identification when character-level accuracy is critical.

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

Immediate Validation and Benchmarking: Once extraction for the current COP has been implemented, the pipeline is tested on the Pilot Corpus and results are compared against manual annotations from Task 3.3. Evaluation strategies range from basic (standard metrics on a pilot corpus) to comprehensive (ablation studies and hybrid approach exploration) based on project constraints. This immediate validation enables rapid identification of extraction bottlenecks or misalignments before proceeding to the next COP.

Output: An extraction pipeline capable of processing raw text and generating structured outputs following the annotation schema for the specific pattern being developed. Only after successfully validating extraction for the current COP should development proceed to the next pattern, returning to Task 3.3 to develop its annotation schema. This cycle continues until extraction pipelines have been developed and validated for all identified COPs using the Pilot Corpus.

3.5 Integration and refinement (Task IV)

Task IV harmonizes the COPs (Task 3.2), annotation schemas (Task 3.3) and pipeline components (Task 3.4) into a coherent end-to-end KE system, transitioning the experimental pipeline to production-ready status. The annotation model emerging from processing individual COPs is consolidated into a production version suitable for full corpus processing.

Pipeline Integration: Modular extraction components developed for individual COPs are integrated into a unified text-to-KG pipeline. Integration addresses dependencies between patterns, ensures consistent entity resolution across components and optimizes overall processing architecture. The integrated pipeline processes documents from raw text to complete KGs, instantiating all identified COPs.

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, interaction effects between COP extractors, inconsistent tool coverage across discourse types and representation generation errors. Testing systematically evaluates performance across document types and semantic phenomena, with particular attention to error propagation through pipeline stages.

Annotation Model Refinement to Production Version: Based on testing results, the annotation model evolves into a production-ready version suitable for both manual annotation and automated extraction. This may involve adding elements crucial for automated extraction – coreference chains spanning multiple COPs, disambiguation tags and confidence indicators – while maintaining backward compatibility with COPs. The production annotation model serves as the schema for GT creation in Task 3.6.

Mapping Algorithm Enhancement: Preliminary mapping algorithms from Task 3.3 are consolidated and enhanced to handle the complete KG structure, improving handling of complex semantic structures emerging from COP interactions 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 KE and evaluation (Task 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 Task 3.5 to test data separate from the Pilot Corpus.

GT Preparation: Comprehensive GT creation using the production annotation model involves annotating a test dataset separate from the pilot corpus, covering all COPs from Task 3.2.

Annotation follows the same paradigms established in Task 3.3, including multiple annotators and inter-annotator agreement measurement when resources permit. The GT serves as the gold standard for systematic evaluation, with mapping algorithms from Task 3.5 applied to generate reference RDF. Test dataset size should balance evaluation rigor with annotation resource constraints, typically ranging from 10 to 50 documents, depending on length and complexity.

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

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

(1) Technical Evaluation: Component-level assessment using precision, recall and F_1-score evaluates individual extraction tasks independently for each COP. Coverage analysis for CQs examines whether the KG contains sufficient information to answer the original CQs.

(2) Semantic Evaluation: KG "rehydration" (Gardent *et al*., 2017; Gangemi *et al*., 2024) enables comparison when structural alignment is impossible. This approach reconstructs natural language text from KG information, using metrics like BLEU (Papineni *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).

(3) Competency-Based Evaluation: SPARQL query suites derived from original CQs verify that the KG satisfies the functional requirements that motivated its construction, aligned with evaluation practices from eXtreme Design (Presutti *et al*., 2009) using tools like TestaLOD (Carriero *et al*., 2019).

Domain Expert Validation: A comprehensive review by domain specialists evaluates extraction quality and coherence, with rehydration enabling evaluation by experts without RDF expertise by presenting KG content as natural language.

Iteration Strategy: Evaluation results may trigger returns to earlier tasks: coverage issues to Task 3.3 or Task 3.5, extraction bottlenecks to Task 3.4, systematic errors requiring architectural restructuring in Task 3.5 or fundamental ontological misalignments necessitating a return to Task 3.2 for COP reassessment.

4. Authenticity assessment case study

This section describes the implementation of the ATR4CH methodology for authenticity assessment debates. The methodology operates on the intersection between ontological representation and textual evidence: the entities, relations and concepts defined by the SEBI ontology must correspond to information that can be identified and extracted from the documents. Following the five-task structure of ATR4CH (Section 3), we present: foundational analysis and design (Section 4.1), iterative annotation schema development (Section 4.2), pipeline architecture development (Section 4.3), system integration and refinement (Section 4.4) and comprehensive evaluation (Section 4.5).

4.1 Foundational analysis and design (Task I)

This subsection implements Task I of the ATR4CH methodology (Section 3.2), establishing a foundational understanding of source materials by analyzing both corpus and ontology to identify COPs for KE. As specified in the methodology, Task I requires three foundational inputs: a corpus of unstructured documents, a target ontology defining KR and CQs specifying information requirements. This subsection presents these inputs and the analytical activities that identify extractable patterns.

4.1.1 Corpus collection and analysis. Following Task I guidelines (Section 3.2), we collected Wikipedia articles on historical forgeries, hoaxes and authenticity controversies through web scraping from Wikipedia's categorical organization system. Initial selection covered 31 categories, including Document and Literary Forgeries, Historical Myths, Conspiracy Theories, Pseudepigraphy and Political forgery. From 1,301 retrieved documents [13], 16 categories and 717 articles were excluded because they presented no scholarly debate or were not about CH items. The final dataset encompasses 581 articles across 15 categories (Table 1).

Articles average 8,150 characters and 1,249 tokens, with unique vocabulary averaging 464 tokens per article, indicating substantial lexical diversity. Figure 3 shows a right-skewed distribution, with most articles ranging from 2k to 15k characters and outliers extending beyond 40k characters.

Distribution across categories (Figure 4) reflects the natural prevalence of different forgery types, with literary forgeries representing the largest category (138 articles), followed by pseudepigraphy forms (136 articles across subcategories), and archaeological and artistic forgeries (132 articles combined).

Token count distribution by category (Figure 5) reveals substantial variability. The Shakespeare authorship question demonstrates the highest token density (nearly 10K tokens), while musical hoaxes and modern pseudepigraphy exhibit more consistent, moderate-length articles.

Notable examples include the *Demodocus* [14], a fabricated Platonic dialogue whose Wikidata entry [15] employs a deprecated rank for authorship attribution to Plato (Figure 6), demonstrating how existing knowledge bases represent disputed attributions.

4.1.2 SEBI ontology analysis. Implementing the ontology analysis component of Task I (Section 3.2), we examined the SEBI ontology to assess which ontological elements can be populated from the source documents. The SEBI ontology [16] (Pasqual, 2025) was developed based on scholarly articles (e.g., Härtel, 2017), a catalogue describing 153 known forgeries from Styria (Haider, 2022), and discussions with an expert diplomatist. The data model represents authenticity assessment claims using RDF-star (Hartig, 2017) as a reification method to represent (possibly concurrent) claim contents and contextual information (Daquino *et al*., 2020).

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
Total	*581*

Source(s): Authors' own work

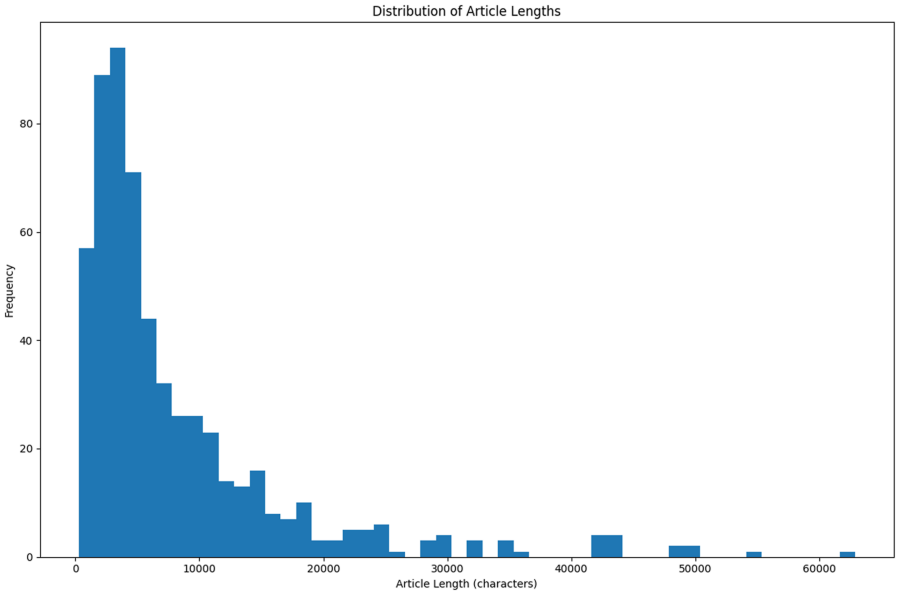

Figure 3. Article length distribution showing a right-skewed pattern characteristic of encyclopedic content. Source: Authors' own work

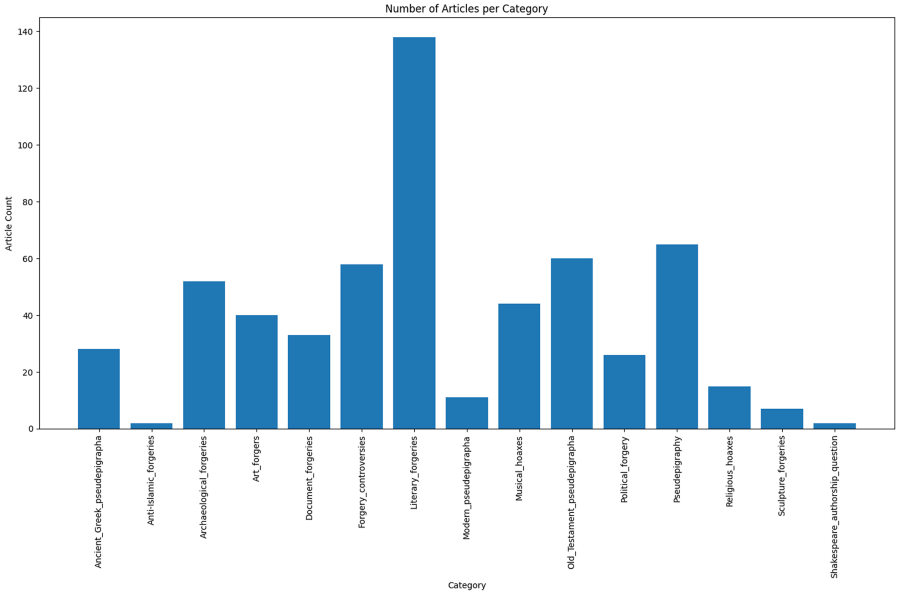

Figure 4. Distribution of articles across Wikipedia categories. Source: Authors' own work

Each claim provides information about the document: authenticity classification, date and place of creation, author, and intention behind creation. Contextual information includes evidence collected by scholars to reach conclusions using evidence-based evaluations, the

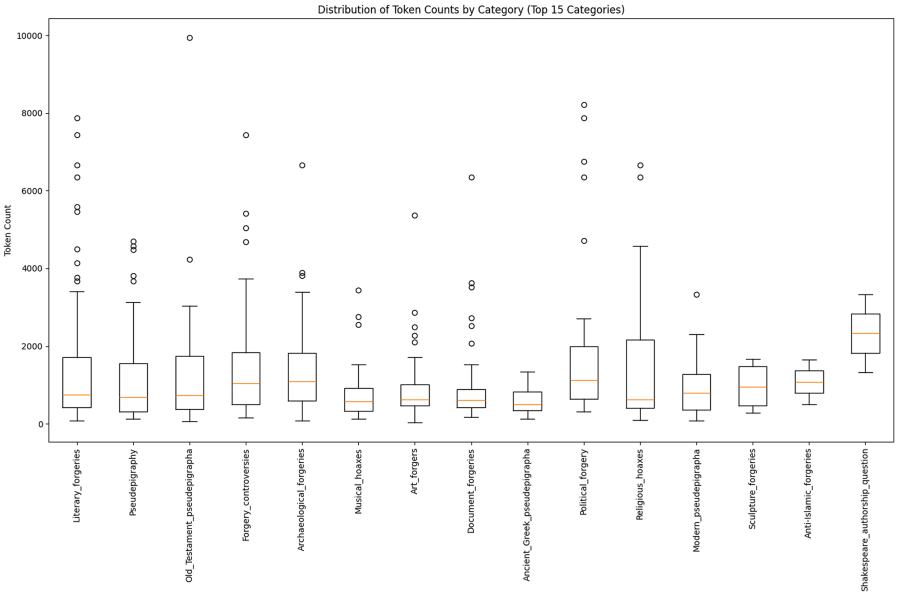

Figure 5. Token count distribution by category showing medians, quartiles and outliers. Source: Authors' own work

Figure 6. Plato noted as the author of the *Demodocus* using a deprecated rank. Source: Wikidata

author of the claim and relevant bibliographic entries (using HiCo [17] and PROV-O [18]). RDF-star (Hartig, 2017) was chosen as the reification method to express both claim content and context, allowing representation of the complete evaluation process.

As shown in Figure 7, each claim contains an authenticity classification. Items are instances of sebi:Forgery, sebi:Authentic, sebi:FormalForgery or sebi:ContentForgery, all subclasses of sebi:Document.

Additionally, each RDF-star quoted triple includes details such as:

(1) the believed creator: sebi:Document → dct:creator → dct:Agent

(2) date of creation: sebi:Document → dct:date → time:Interval

(3) location of creation: sebi:Document → dct:coverage → dct:Location

(4) intention behind creation: sebi:Document → sebi:intended → sebi:Intention.

The dct:date property connects to time:Interval, which includes time:hasBeginning and time:hasEnd properties to specify creation periods and handle fuzzy time-spans. Concerning contextual information (Figure 8), each interpretation (set of claims represented as quoted triples) is categorized as hico:InterpretationAct connected to prov:Agent for authoriality and linked to supporting evidence (sebi:support sebi:Evidence).

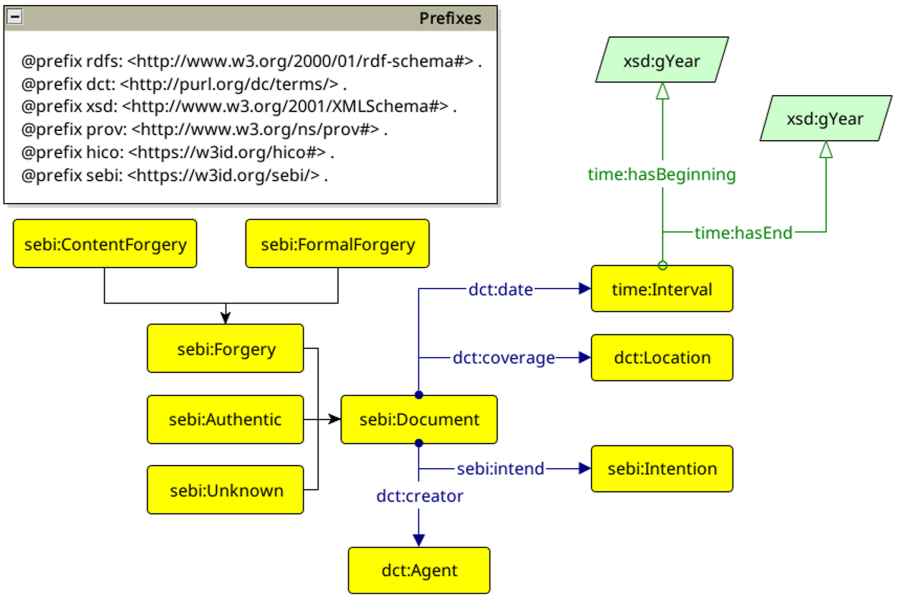

Figure 7. Selection of classes and properties to represent scholarly claims addressing authenticity assessment of a document. Source: Authors' own work

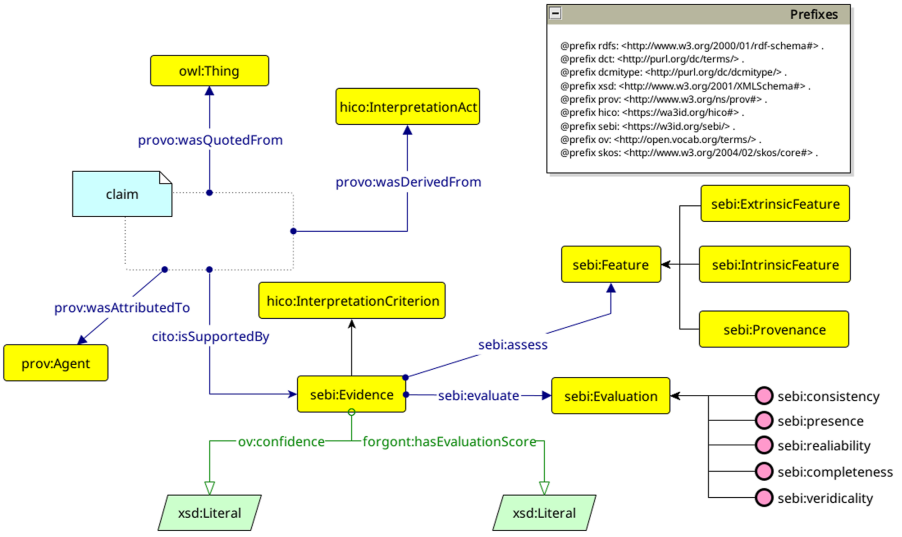

Figure 8. Selection of classes and properties to represent contextual information about scholarly claims addressing authenticity assessment of a document. Source: Authors' own work

Document features and their evaluation are components of the ontology. Document features (sebi:Feature) are extrinsic features (sebi:ExtrinsicFeature), intrinsic ones (sebi:IntrinsicFeature) or provenance information (sebi:Provenance), capturing aspects such as ink, support, handwriting and orthography. Each feature is evaluated on established criteria (sebi:Evidence) such as consistency, presence, completeness, veridicality and reliability. A score is associated with each

evidence as xsd:Literal using forgont:hasEvaluationScore. The evaluation score allows integration of negatives (e.g., absence of signature is represented as evidence based on the feature "authentication marks" with evaluation "presence" and score false or 0).

4.1.3 COPs identification. We identified COPs following the COP identification process specified in Task I (Section 3.2). The identification process involved: (1) assessing alignment between CQ, ontological structures and available textual content, (2) identifying patterns with sufficient textual evidence for reliable extraction, (3) prioritizing based on extractability feasibility and CQ relevance and (4) selecting a manageable subset forming the semantic backbone for KE.

Four COPs were identified for extraction, presented in hierarchical priority order:

(1) CH Item Metadata – Alleged information that the object claims about itself (creator, date, location) before scholarly critical analysis. For instance, in the case of the Donation of Constantine, a relevant CQ would be "What does the Donation claim about its author, date and place of creation?"

(2) Scholarly Opinions – Authenticity assessments expressed by scholarly agents, classified as Authentic, Forgery, FormalForgery or ContentForgery. A typical pattern would be "Scholar X evaluates Document Z and concludes Authenticity Status S." For instance, "Which scholars identify the Donation of Constantine as authentic?"

(3) Evidential Features – Characteristics examined by scholars to support assessments, organized by type with evaluations. Example CQ: "What evidence does Lorenzo Valla cite to support his assessment of the Donation?"

(4) Alternative Hypotheses – Competing scholarly claims about the actual creator, date, location or intended purpose. Example CQ: "What are the competing scholarly hypotheses about the actual date of the Donation's creation?"

These COPs emerged from the intersection of CQs, ontological structures and extractable content patterns identified in corpus analysis, following the systematic approach described in the methodology.

4.1.4 Pilot corpus selection. Following the Task I guidelines for pilot corpus selection (Section 3.2), we defined a qualitative sample of the corpus. Seven articles were chosen (*Donation of Constantine*, *Eremin Letter*, *Getty Kouros*, *Historia Augusta*, *Life of Homer*, *Marriage Charter of Empress Theophanu* and *Protocols of the Elders of Sion*), each belonging to a different category. Selection criteria included: (1) presence of multiple scholarly perspectives, (2) clear attribution of claims, (3) discussion of evidence-based reasoning and (4) representation of different temporal periods and document types.

4.2 Minimal working annotation development (Task II)

This subsection documents the implementation of Task II (Section 3.3), developing annotation schemas incrementally by processing one pattern at a time. Each pattern identified in Task I received iterative schema development, pilot corpus application and RDF mapping validation before proceeding to the next pattern.

Annotation Schema Design. Following Task II guidelines, the annotation model was developed through INCEpTION (Klie *et al.*, 2018), implementing three core patterns from SEBI corresponding to the first two COPs: CH item metadata, scholarly agents and authenticity opinions. CH item metadata establishes entity linking using INCEpTION's Wikidata integration, reconciling item types to DCMI Type Vocabulary classes (dcmitype:Text, dcmitype:PhysicalObject and dcmitype:Collection). Scholarly agents (Cognizers) correspond to dct:Agent, linked to Wikidata when possible, with fallback strategies for entities without entries. Authenticity claims were modeled through directed relations between Cognizer and item spans, labeled according to SEBI's authenticity categories (Authentic, FormalForgery, ContentForgery, Forgery and Neutral) as shown in Figure 9.

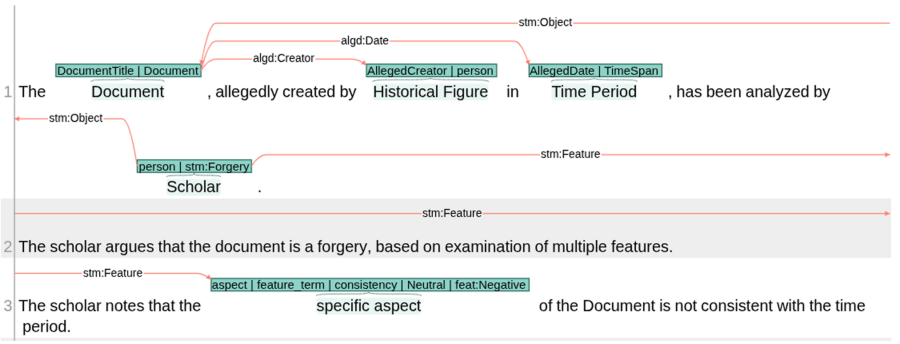

Figure 9. Example annotation of an entity expressing an opinion about a CH item. Source: Authors' own work

Knowledge Base Integration. Integration with Wikidata through INCEpTION's Knowledge Base module ensures consistent entity identification, preventing inconsistent annotation that would hamper aggregation and reasoning in the final KG. Wikidata was selected for its domain coverage, data quality and compatibility with the entity types required by the COPs.

Annotation Process. Following established corpus linguistics and NLP practices, a single experienced annotator performed the annotation work iteratively, with guidelines refined based on encountered edge cases and thoroughly documented to ensure reproducibility.

RDF Mapping Validation. Preliminary mapping exercises validated that the annotation schema could produce target knowledge structures. Using the Pilot Corpus, annotation-to-RDF mapping was validated through the algorithm in Listing 4.2, serving as a unit test confirming that annotation patterns correctly transformed to valid RDF instantiating SEBI classes and properties.

```
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 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}.
```

Successful RDF generation from annotations validated the initial schema design, confirming that annotation patterns adequately captured the semantic content required to instantiate SEBI's authenticity claim structure. Following successful validation, development proceeded to the next pattern following the iterative cycle specified in Task II.

4.3 Pipeline architecture development (Task III)

This subsection implements Task III of the ATR4CH methodology (Section 3.4), designing and implementing computational tools to automatically extract COPs from text using the annotation model developed in Task II as the target schema. Following the methodology's emphasis on incremental development, extraction capabilities were developed and validated for each COP using the same Pilot Corpus documents that were annotated in Task II, rather than building the complete extraction pipeline before testing.

4.3.1 Architectural design and tool selection. Following the task decomposition and tool selection guidelines in Task III (Section 3.4), the pipeline was designed around the annotation model elements developed in Task II, prioritizing based on COP semantic importance and accounting for information manifestation patterns identified during corpus analysis in Task I. The KE task addresses CH corpora's domain-specific characteristics and limited annotated training data through a modular architecture enabling incremental extraction.

The pipeline integrates three complementary technologies, each addressing specific extraction requirements identified through the methodology:

GliNER for NER: Provides lightweight, generalist NER using custom entity types, enabling precise character-level span identification for entity extraction without requiring task-specific training data. This addresses the limited annotated training data characteristic of CH domains, as identified in the methodology's tool selection strategy.

LLMs for Structured Extraction: Handle complex information extraction through JSON schema-based responses (Schick *et al*., 2023; Qin *et al*., 2024) using ICL strategies (Brown *et al*., 2020; Min *et al*., 2022). We evaluated three models at varying parameter scales to understand performance trade-offs and cost-effectiveness, following the methodology's resource-adaptive approach:

(1) Claude Sonnet 3.7 [19] as the largest model

(2) Llama 3.3 70B (Dubey *et al*., 2024) as a medium-sized model

(3) GPT-4o-mini [20] as the smallest model [estimated 8–14 billion active parameters (Ben Abacha *et al*., 2025)].

Rule-Based Entity Linking: Employs the Wikibase API [21] with domain-specific heuristics, integrating the knowledge base resources identified during annotation model development (Task II, Section 4.2). After evaluating 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.

Paragraph-Level Processing Strategy: While the selected LLMs can process complete documents, the system automatically selects only relevant paragraphs whenever possible. This design serves three purposes: (1) reducing content volume per processing step to minimize potential opinion overlap between entities, which we hypothesize improves precision; (2) demonstrating scalability to documents of arbitrary length; and (3) maintaining computational efficiency and cost-effectiveness by minimizing token consumption per API call.

4.3.2 Sequential pipeline implementation. Following the modular approach specified in Task III (Section 3.4), the KE pipeline consists of six sequential components, each enriching the output before passing it to the next. Each component targets specific annotation schema elements from Task II and produces a JSON output following a predefined schema designed for conversion to RDF. Development proceeded through preliminary implementations and sequential testing over the COPs identified in Task I until the complete KG could be extracted.

Figure 10 presents the pipeline architecture, showing the correspondence between pipeline components and the COPs identified in Task I:

(1) Metadata Extraction: Targets COP 1 (CH Item Metadata) – Raw text documents → alleged and settled item metadata

(2) Opinion Holder Identification: Targets COP 2 (Scholarly Opinions) – Item metadata + text → entity mentions with opinion classifications

(3) Entity Resolution: Integrates knowledge bases identified in Task II – Entity mentions → Wikidata-linked entity clusters

(4) Opinion Extraction: Completes COP 2 extraction – Linked entities + paragraphs → structured authenticity opinions

(5) Evidence Mining: Targets COP 3 (Evidential Features) – Opinions + contexts → feature evaluations with polarity

(6) Hypothesis Extraction: Targets COP 4 (Alternative Hypotheses) – Evidence + full context → conflicting statements and alternative theories.

The following subsections detail each component's implementation, demonstrating how the annotation schemas developed in Task II guided extraction design.

4.3.3 Component 1: CH item metadata extraction.

COP Addressed: COP 1 (CH Item Metadata, Section 4.1.3)

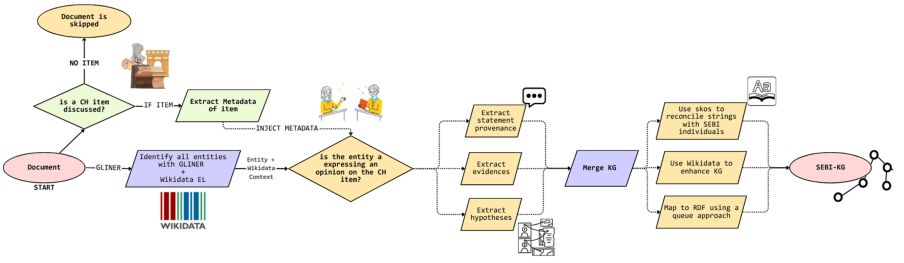

Figure 10. Sequential pipeline architecture for SEBI-based KG generation, showing correspondence between components and COPs. Source: Authors' own work

Annotation Schema: CH Item Metadata Layer (Section 4.2)
Input: Raw Wikipedia articles in markup (.txt files)
Output: Cleaned articles; JSON with alleged item metadata.

This component implements extraction for the first COP, identifying and extracting metadata about CH items discussed in each article. The LLM is instructed to extract a JSON schema describing all items under discussion, specifically targeting the *alleged metadata* elements defined in the annotation model – what items claim to be – including purported authors, creation dates, locations, item types and subject matter. The task relies on ICL in a Few-Shot setting (with three examples), using CoT reasoning. Figure 11 shows an example input and JSON output.

4.3.4 Component 2: cognizer identification.

COP Addressed: COP 2 (Scholarly Opinions, Section 4.1.3)
Annotation Schema: Entity and Opinion layers (Section 4.2)
Input: Cleaned article + item metadata (output of Component 1)
Output: JSON with is_cognizer classification, coreferences.

This component begins extraction of the second COP by identifying scholarly agents. Following the entity identification approach developed during annotation, it employs GliNER for NER, targeting people, organizations, groups and locations as specified in the annotation schema. GliNER identifies precise character-level spans, enabling exact identification and grouping of paragraphs in which each entity appears. The LLM then performs binary classification (is_expressing_opinion: True/False) to identify which entities function as Cognizers, alongside additional textual mentions and co-references. The task relies on ICL in a Few-Shot setting (with three examples), using CoT reasoning.

4.3.5 Component 3: entity resolution and linking.

COP Addressed: Supporting infrastructure for all COPs
Annotation Schema: Knowledge Base integration (Section 4.2)
Input: Cognizers and coreferences (output of Component 2)
Output: JSON with relevant paragraphs grouped by Cognizer and biographical information.

This component implements the knowledge base integration strategy established during annotation model development (Task II, Section 4.2). It performs coreference resolution and Entity Linking through a three-stage pipeline. First, it collects all entity mentions across paragraphs and groups them by exact string match. Second, it applies coreference resolution by calculating mention similarity using a Jaccard coefficient over word sets (with common stop words removed), clustering mentions with similarity scores exceeding 0.7 when entity types are compatible. The system selects the longest string as the representative mention for each cluster. Third, it performs Entity Linking by querying the Wikidata Search API (wbsearchentities) with each mention variant, retrieving up to five candidates per query and deduplicating results by Wikidata identifier.

Each candidate undergoes scoring through the Wikidata API (wbgetentities) to retrieve claims and labels. The scoring function combines three weighted components: (1) name

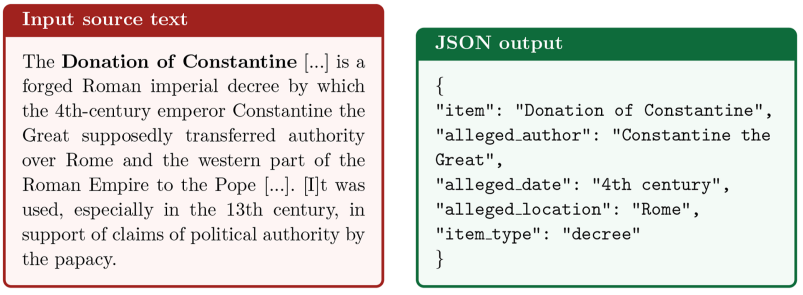

Figure 11. CH item metadata extraction from source text to structured JSON output

similarity between cluster mentions and candidate labels/aliases, calculated using Jaccard similarity over word sets with special handling for first name/initial matching (weight: 0.6 for labels, 0.3 for aliases); (2) entity type compatibility verified through Wikidata property P31 (instance of), with type mappings defined for persons (Q5), organizations (Q43229, Q7278), locations (Q2221906) and groups (Q16334295); and (3) for person entities, occupation relevance assessed through Wikidata property P106 (occupation), comparing retrieved occupation identifiers against a curated vocabulary of scholarly occupations (weight: 0.1). The system applies a minimum threshold of 0.4 for candidate acceptance and selects the highest-scoring candidate per cluster. When multiple candidates exceed the threshold, the system retrieves detailed data in batch requests and calculates an overlap score measuring the proportion of cluster mentions matching each candidate's labels and aliases (using both exact matches and fuzzy matching with a similarity threshold of 0.7), combining this with the initial score through weighted average (0.6 for overlap, 0.4 for initial score).

4.3.6 Component 4: opinion extraction and classification.

COP Addressed: COP 2 (Scholarly Opinions, Section 4.1.3)

Annotation Schema: Opinion classification layer (Section 4.2)

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

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

This component completes extraction of the second COP by extracting and classifying authenticity opinions. If the entity has been successfully linked to Wikidata in Component 3, this information is provided to the model. The extraction process captures the main elements of the annotation schema: opinion targets (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.

4.3.7 Component 5: evidence mining and feature assessment.

COP Addressed: COP 3 (Evidential Features, Section 4.1.3)

Annotation Schema: Evidence and Features Layer (Section 4.4.3)

Input: Structured opinions (output of Component 4) + contextual paragraphs

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

This component implements extraction for the third COP, enriching opinions with evidences and features being evaluated. Features are organized into three categories following the SEBI ontology and the annotation model developed in Task II: *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, presence, completeness, reliability and veridicality, as specified in the annotation schema. Each evaluation receives polarity assignment (positive, negative, neutral evidence) and links to supporting scholarly opinions.

4.3.8 Component 6: hypothesis extraction.

COP Addressed: COP 4 (Alternative Hypotheses, Section 4.1.3)

Annotation Schema: Scholarly Hypotheses Layer (Section 4.4.3)

Input: Opinions + evidence evaluations + full document context

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

This final component implements extraction for the fourth COP, enriching the output with scholars' hypotheses. The hypotheses types correspond directly to the relation types defined in the annotation model: *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. For consistency and to avoid negated categories, polarity (positive/negative) is included as a field.

4.3.9 Immediate validation on pilot corpus. Following the immediate validation strategy specified in Task III (Section 3.4), once extraction for each COP was implemented, the pipeline was tested on the Pilot Corpus and results were compared against manual annotations created in Task II. This immediate validation cycle enabled rapid identification of extraction bottlenecks or misalignments before proceeding to the next COP. This cycle continued until extraction pipelines had been developed and validated for all identified COPs using the Pilot Corpus.

4.4 Integration and refinement (Task IV)

This subsection implements Task IV of the ATR4CH methodology (Section 3.5), harmonizing the COPs (Task I), annotation schemas (Task II) and pipeline components (Task III) into a coherent, end-to-end KE system. Task IV represents a critical transition point: the modular components developed and validated on the Pilot Corpus are now integrated into a production-ready system, and the annotation model that emerged from processing individual COPs is consolidated into a production version suitable for full corpus processing and comprehensive GT creation.

4.4.1 Pipeline integration. Following the integration approach specified in Task IV (Section 3.5), the modular extraction components developed for individual COPs were integrated into a unified text-to-knowledge-graph pipeline. Integration addressed dependencies between patterns, ensured consistent entity resolution across components and optimized the overall processing architecture. The integrated pipeline processes documents from raw text to complete KGs that instantiate all identified COPs.

4.4.2 End-to-end pipeline testing. Following the comprehensive testing approach in Task IV (Section 3.5), testing over the pilot corpus processed documents from raw text to final KGs, revealing systematic issues including data sparseness patterns, interaction effects between COP extractors, inconsistent tool coverage across discourse types and representation generation errors. Testing systematically evaluated performance across document types and semantic phenomena, with particular attention to error propagation through pipeline stages.

4.4.3 Production annotation model development. Following the annotation model refinement approach in Task IV (Section 3.5), based on testing results, the annotation model evolved into a production-ready version suitable for both manual annotation and automated extraction. Following successful development and validation of extraction pipelines for individual COPs (Tasks II–III), the annotation schemas were consolidated into a production model suitable for comprehensive GT creation. This refined model captures CH item metadata, evidence and features, and scholarly hypotheses through additional layers developed iteratively during pipeline testing on the Pilot Corpus, maintaining backward compatibility with COPs while adding elements crucial for automated extraction.

CH Item Metadata Layer. This layer captures *alleged metadata* – descriptive information (creator, date, location) that the document or artifact purports about itself, before scholarly critical analysis. This includes face-value claims presented within the item or by whoever claimed to find the item regarding authorship, creation date, geographic origin and other identifying characteristics. Annotations include AllegedCreator, AllegedDate, AllegedLocation, ItemSubject and ItemType, plus properties for formal forgeries (ItemCreator, ItemDate, ItemLocation) (see Figure 12).

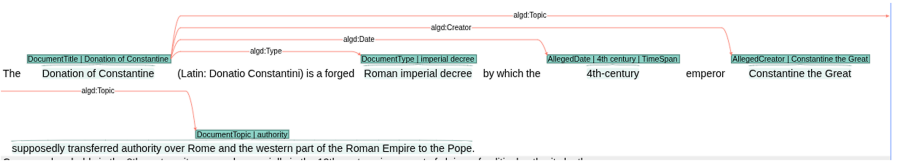

Figure 12. Alleged metadata annotation for the *Donation of Constantine*. Source: Authors' own work

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

Consider Lorenzo Valla's assessment of the Donation's language features (Figure 15), which converts to three evidence structures linking textual features to evaluation criteria and polarities.

Listing 4.4.3 shows the evidence mapping algorithm.

```
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 (see Figures 13–15).

Listing 4.4.3 details the hypotheses mapping algorithm.


```
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:{unifiedTargetSpan} 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:wasDerived From kb:cognizer_uri_about_item_uri.
```

4.4.4 Mapping algorithm enhancement. Following the mapping algorithm enhancement approach in Task IV (Section 3.5), preliminary mapping algorithms from Task II were consolidated and enhanced to handle the complete KG structure. As the output JSON model closely resembles the GT structure, the mapping uses similar logic, differing only in that it

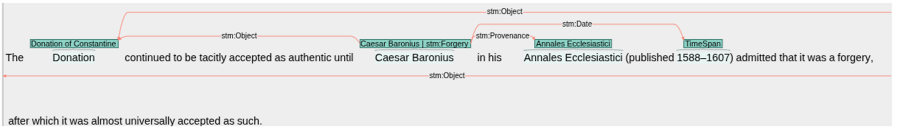

Figure 13. Caesar Baronius's admission of forgery with provenance annotation. Source: Authors' own work

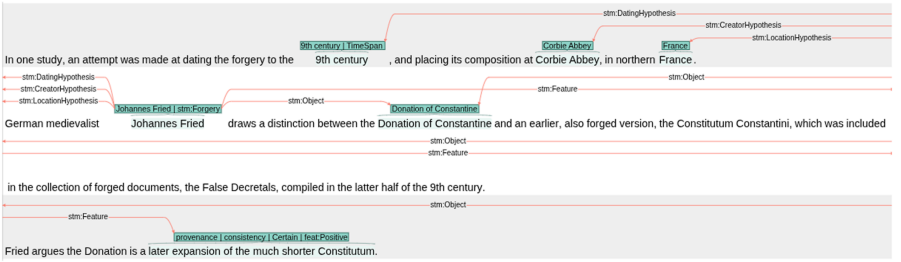

Figure 14. Johannes Fried's hypotheses annotation for the *Donation of Constantine*. Source: Authors' own work

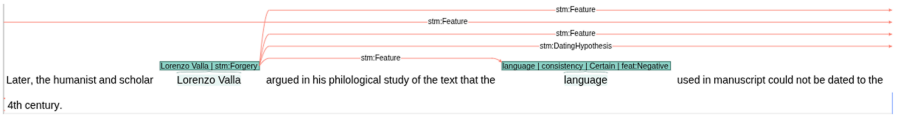

Figure 15. Lorenzo Valla's opinion with feature assessment annotation. Source: Authors' own work

processes from JSON rather than the JSON UIMA CAS (Content Analysis System) format used by INCEpTION. Error handling mechanisms manage extraction failures and partial results using validation tools such as SHACL and OWL reasoners as specified in the methodology.

4.4.5 GT statistics. Each annotation layer maps to RDF following SEBI ontology principles, with Wikidata integration providing entity resolution. The INCEpTION project is available on GitHub alongside mapping scripts [22]. Statistics for annotation results are shown in Table 2.

4.5 KE and evaluation (Task V)

This subsection implements Task V of the ATR4CH methodology (Section 3.6), employing technical validation and domain-expert evaluation to ensure knowledge structures accurately represent domain-specific discourse complexity. The refined system from Task IV is applied to

Table 2. GT annotation results

Span	Count
CH items	45
Entities	235
Interpretation acts	215
Evidences	132
Features	115
Wikidata alignments	308

Source(s): Authors' own work

test data separate from the Pilot Corpus used for development, representing the first application of the complete pipeline to unseen documents. The final iterations of the prompts are available in Appendix.

4.5.1 KG generation. Following the KE approach in Task V (Section 3.6), test datasets were processed through the complete pipeline under realistic deployment conditions, with systematic documentation of performance and failure modes. This represented the first application of the extraction pipeline beyond the Pilot Corpus used for development. The final output is mapped to RDF using the algorithms explained in previous sections. This subsection showcases produced KGs in RDF-star format (specifically, this example was generated by the pipeline using Llama 3.3 70B).

Figure 16 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 claims to be) and scholarly assessments, as shown in Listing 4.5.1. The Donation of Constantine exemplifies this pattern:

Document representation with alleged and scholarly metadata

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

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

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

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

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

Listing 4.5.1 shows Lorenzo Valla’s interpretation of the Donation.

Lorenzo Valla’s interpretation with supporting evidence

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

# Valla’s interpretation act
kb:lorenzo_valla_about_donation_of_constantine a hico:InterpretationAct;
     sebi:date kb:1,439–1,440;
     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 4.5.1.

```
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.
```

4.5.2 Evaluation framework. Following the multi-level evaluation approach specified in Task V (Section 3.6), the evaluation framework provides a multi-dimensional assessment of the KG generation pipeline. Multiple complementary approaches address KG evaluation challenges, integrating human assessment throughout the evaluation pipeline and using F_1 score and G-EVAL metrics. The framework systematically addresses five EQs, based on the CQs that defined the original SEBI ontology (Pasqual, 2025).

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

Methodology: This is formulated as a multiclass classification task, evaluating the metadata extraction component against the GT. The classification scheme follows standard evaluation practices: True Positive (TP) for exact matches between model output and GT, False Positive (FP) for incorrect model predictions, True Negative (TN) for correctly identified absence of metadata when GT is also empty and False Negative (FN) for missing outputs when GT contains valid metadata. To accommodate acceptable semantic variations (e.g., alternative titles, location aliases), all FP cases are manually reviewed to identify outputs that are semantically equivalent to the GT and should be reclassified as TP.

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

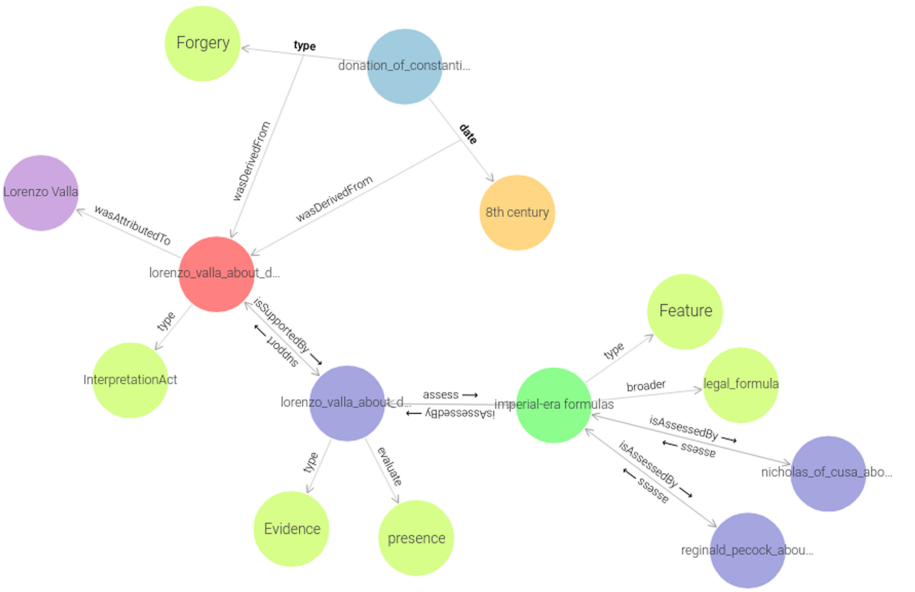

Figure 16. Lorenzo Valla's statement about the *Donation of Constantine*. Source: Authors' own work

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: The entity extraction component is evaluated by conducting frequency-based analysis comparing GT entities with model-identified entities.

Metrics: Entity-level recall (proportion of GT entities correctly identified) and 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 the ontological framework, where each piece of evidence comprises multiple semantic dimensions (evaluated feature, evaluation perspective, broader feature class, polarity), a custom scoring metric operating on a 4-point scale is implemented. For each evidence prediction, points are assigned 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.

Score Interpretation: 0 points indicates complete extraction failure (equivalent to FN or total FP); 1–2 points indicates weak but partially acceptable outputs; and 3–4 points indicates acceptable to strong outputs meeting semantic requirements.

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

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

Methodology: The same precision, recall and F_1-score evaluation framework established for EQ1 is applied to assess the hypothesis extraction component. Model outputs are compared

against expert-annotated GT for both specific scholarly hypotheses and overall authenticity determinations.

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

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

Methodology: To evaluate representation fidelity, G-EVAL (Liu *et al*., 2023) is employed. Since the KGs only represent opinions inside the text, comparing the source document with a rehydrated version of the KG would heavily bias the evaluation metric. This led to avoiding similarity-based metrics like BLEU, ROUGE and COMET with the source corpus as used in Gangemi *et al*. (2024).

G-EVAL evaluates 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.

Previous evaluation metrics mostly covered matchable entries between GT and output, whereas G-EVAL evaluates the complete output.

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

5. Results

This section presents a comprehensive evaluation and preliminary discussion of findings across the five EQs outlined in Section 4.5. 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 and Llama for brevity). We begin with simple exploratory SPARQL queries across the 3 KGs and compare the results with the GT, as shown in Figure 17.

Table 3 and Figure 18 provide an overview of the KGs generated by each model compared to the GT. The models produce more triples than the GT (10,000–12,000 vs 4,026), as 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, the models generate comparable numbers of Interpretation Acts and Cognizers to the GT, suggesting at this stage a similar density of extracted information.

5.1 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_1-scores ranging from 0.97 to 0.99.

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"))
}
```

Figure 17. SPARQL query used to extract entity counts from the KGs for statistical comparison across models

Table 3. KE overall metrics

Model	Triples	Interpretation acts	Cognizers
GT	4,026	170	164
Claude	10,173	148	103
GPT	12,088	247	201
Llama	10,119	217	172

Source(s): Authors' own work

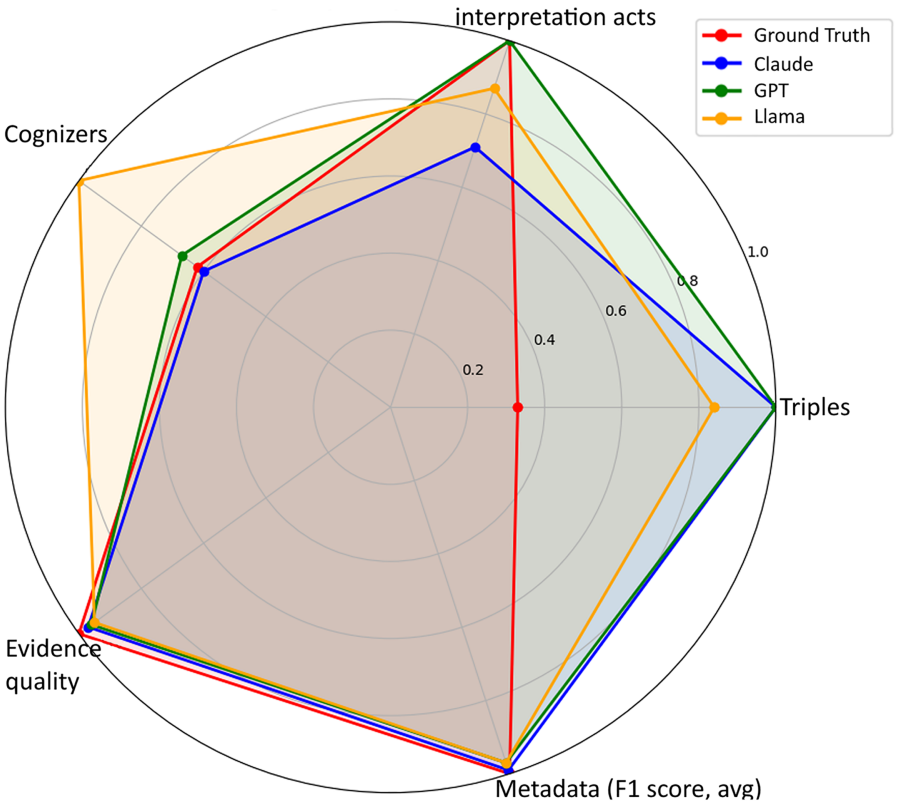

Figure 18. Radar chart of different KG extractions. Source: Authors' own work

Claude Sonnet 3.7 achieves the highest overall performance with an F_1-score of 0.987. All models show nearly perfect recall, indicating successful extraction of all relevant metadata elements, with precision differences primarily reflecting varying FP rates. Date extraction shows more variability, with Llama 3.3 achieving the lowest precision (0.867) due to higher FP 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 sizes.

Table 4. CH item metadata extraction performance across three LLMs

Model	Category	Precision	Recall	F_1-score
Claude	Titles	1.000	1.000	1.000
	Type	1.000	1.000	1.000
	Creators	0.977	0.977	0.977
	Dates	0.978	1.000	0.989
	Locations	0.978	1.000	0.989
	Overall	*0.987*	*0.995*	*0.991*
GPT	Titles	0.889	1.000	0.941
	Type	0.956	1.000	0.977
	Creators	0.956	1.000	0.977
	Dates	0.911	1.000	0.953
	Locations	1.000	1.000	1.000
	Overall	*0.942*	*1.000*	*0.970*
Llama	Titles	0.933	1.000	0.966
	Type	0.933	1.000	0.966
	Creators	0.933	1.000	0.966
	Dates	0.867	1.000	0.929
	Locations	1.000	1.000	1.000
	Overall	*0.933*	*1.000*	*0.965*

Source(s): Authors' own work

5.2 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 that 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%.

5.3 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

Table 5. Entity recognition coverage and accuracy

Model	Precision	Recall	F_1	TP	FP	FN
Claude	0.696	0.763	0.728	71	31	22
GPT	*0.718*	*0.912*	*0.803*	145	57	14
Llama	0.626	0.817	0.709	107	64	24

Source(s): Authors' own work

Table 6. Evidence extraction quality and coverage

Model	Mean score (0–4)	Percentage score (%)
Claude	*3.87*	*96.8*
GPT-4o-mini	3.84	96.0
Llama 3.3	3.81	95.3

Source(s): Authors' own work

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.

5.4 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 authenticity judgments.

GPT-4o-mini achieves the highest overall F_1-score (0.749) for hypothesis extraction, with particularly strong performance in authenticity type classification (0.845). However, the model shows weaker performance in creator hypothesis identification (0.484), suggesting challenges in extracting attribution hypotheses. Claude demonstrates exceptional performance in geographic hypotheses (0.923 F_1) and temporal hypotheses (0.791 F_1), indicating strength in extracting location and dating alternative theories. Llama 3.3 shows the most balanced performance across hypothesis types, with particularly strong creator hypothesis extraction (0.712 F_1). The variation across hypothesis types reflects the inherent complexity of scholarly reasoning, with location and date hypotheses generally more explicitly stated than creator attributions or underlying motivations.

5.5 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* and *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. Table 8 shows the per-statement G-EVAL scores for the three models, while Table 9 for the whole KG produced by a single Wikipedia article. This evaluation confirms a key pattern in our pipeline – when an entity is correctly identified as a

Table 7. Hypothesis and judgment extraction performance

Model	Macro F_1	Type F_1	Creator F_1	Date F_1	Location F_1
Claude	0.655	0.652	0.638	*0.791*	*0.923*
GPT	*0.749*	*0.845*	0.484	0.595	0.727
Llama	0.694	0.691	*0.712*	0.762	0.727

Source(s): Authors' own work

Table 8. Per-statement correctness (G-EVAL scores on 0–1 scale)

Model	Mean	Std dev	Range
Claude	*0.620*	0.133	0.333–0.889
GPT	0.590	0.204	0.222–0.889
Llama	0.533	0.153	0.222–0.889

Source(s): Authors' own work

Table 9. Overall debate representativeness (G-EVAL scores on 0–1 scale)

Model	Mean	Std dev	Range
Claude	*0.607*	0.121	0.333–0.889
GPT	0.580	0.199	0.222–0.889
Llama	0.523	0.144	0.222–0.778

Source(s): Authors' own work

Cognizer, its 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.

6. 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.

6.1 Methodological framework validation

To answer RQ1, our five-step ATR4CH methodology proves effective in developing the pipeline within the boundaries of an ontology. The granular evaluation demonstrates that our *divide-and-conquer* methodology enables systematic refinement of individual components while maintaining system coherence. The result is 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.

6.2 Extraction performance analysis

To answer RQ2, our evaluation reveals component-specific performance patterns across all tested models. Performance varies 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). Extracting alleged metadata proves straightforward across models, while capturing nuanced scholarly hypotheses requires more sophisticated interpretation regardless of architecture. The evidence extraction results demonstrate that contemporary LLMs can effectively capture

multi-dimensional evidential reasoning, but they can do so only *when they can identify the Cognizer* – this represents an error propagation problem we identified in the pipeline, as the out-of-GT outputs for evidence extraction are mostly empty or incorrect.

6.3 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.

6.4 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. 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.

6.5 Deployment implications and cost-effectiveness

To answer RQ5, the performance differences between models are relatively modest, while model sizes and costs differ substantially [23]. 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.

The methodology's adaptability accommodates diverse institutional landscapes: smaller projects can benefit from intensive human-in-the-loop approaches with API-based models, while larger projects can leverage automated scaling through extensive annotation datasets and local deployment.

6.6 Contributions, limitations and future directions

In this work, 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 integration 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.

AI tool disclosure

This research employed LLMs (Claude Sonnet 3.7, Llama 3.3 70B and GPT-4o-mini) as research subjects for KE 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. Claude Sonnet 4.5 was used as a writing aid for summarization, translation and code generation when necessary. No AI tools were used in the ideation, analysis or interpretation of this manuscript beyond the specified cases above. The authors maintain full responsibility for the research design, methodology, data interpretation, figure and listing creation, and all conclusions presented.

Acknowledgments

This research was partially funded by the European Union – Next Generation EU, investment I.4.1 PNRR Patrimonio Culturale, Decreto Ministeriale *n.* 351, 9 April 2022. We gratefully acknowledge Dr. Cristina Solidoro (Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy) for her bibliographic suggestions and additional domain expertise in support of the case study.

(The Appendix follows overleaf)

Appendix

Metadata Extraction System Prompt

You are an expert Knowledge Extraction agent. Your task is to extract factual claims about cultural heritage documents - distinguishing between what documents claim about themselves versus what scholars believe about them.

Task

Extract alleged metadata from Wikipedia-style text:
- Alleged authorship: Who the document claims to be by
- Alleged dating: When the document claims to be from
- Alleged location: Where the document claims to originate
- Publisher/Discoverer: Who published, discovered, or brought the document to light
- Actual authorship: Who scholars think really created it
- Actual dating: When scholars think it was really created
- Actual location: Where scholars think it really originated

Examples

Example 1: Chronicle of Valdoria

Input:

The Chronicle of Valdoria was published in 1962 by antiquarian dealer Giuseppe Torretti, who claimed to have discovered the manuscript in the archives of San Pietro monastery. The document purports to be a 12th-century chronicle written by Brother Marcus documenting the founding of the monastery in northern Italy. However, paleographic analysis conducted by Dr. Elena Rossi in 2018 revealed that the parchment contains watermarks not used until the 15th century, and the Latin contains grammatical constructions typical of Renaissance humanists rather than medieval scribes.

Output:

```
{
    “documents”: [
      {
        “document”: “Chronicle of Valdoria”,
        “alleged_metadata”: {
          “alleged_author”: [“Brother Marcus”],
          “alleged_date”: “12th century”,
          “alleged_location”: “northern Italy”,
          “publisher”: [“Giuseppe Torretti”],
          “actual_author”:””“,
          “actual_date”: “15th century”,
          “actual_location”:””
          }
      }
    ]
}
```

Example 2: Codex Aureus Britannicus

Input:

In 1924, German manuscript dealer Heinrich Weber announced the discovery of the Codex Aureus Britannicus, which he claimed to have acquired from a private English collection. The manuscript purports to be an illuminated Gospel book created by Celtic monks at Iona Abbey in the 8th century, allegedly commissioned by King Aethelbald of Mercia. The codex bears an inscription stating it was “written in the year of our Lord 742 by the hand of Brother Columba.” Modern forensic analysis by

↪Cambridge University, however, determined that the gold leaf contains titanium
↪dioxide, a pigment not available until 1916, suggesting the manuscript was created
↪in the early 20th century, possibly in Germany.

Output:

```
{
    "documents": [
        {
            "document": "Codex Aureus Britannicus",
            "alleged_metadata": {
                "alleged_author": ["Brother Columba"],
                "alleged_date": "742",
                "alleged_location": "Iona Abbey",
                "publisher": ["Heinrich Weber"],
                "actual_author":"",
                "actual_date": "early 20th century",
                "actual_location": "Germany"
            }
        }
    ]
}
```

Example 3: Letters of Empress Theodora

Input:

The Letters of Empress Theodora were first published by Turkish historian Mehmet Ozkan in
↪1925, who claimed the documents were discovered during restoration work at Hagia
↪Sophia in Constantinople. The letters allegedly consist of correspondence between
↪the Byzantine Empress Theodora and various nobles, purporting to reveal court
↪intrigues in 6th-century Constantinople. Professor Andreas Mikhailov's textual
↪analysis, published in 2019, demonstrated that the Greek contains modern
↪grammatical forms and references to concepts unknown in the Byzantine period,
↪concluding the letters were fabricated in the 1920s by an unknown forger seeking to
↪capitalize on interest in Byzantine history.

Output:

```
{
    "documents": [
        {
            "document": "Letters of Empress Theodora",
            "alleged_metadata":{
                "alleged_author": ["Empress Theodora"],
                "alleged_date": "6th century",
                "alleged_location": "Constantinople",
                "publisher": ["Mehmet Ozkan"],
                "actual_author": ["unknown"],
                "actual_date": "1920s",
                "actual_location":""
                }
        }
    ]
}
```

Example 4: Multiple Documents – Venetian Statue Collection

Input:

In 1889, art dealer Rodolfo Marinetti claimed to have acquired three ancient Roman marble
↪statues from excavations near Venice. The Venus of San Marco purports to be a 1st-
↪century sculpture by the workshop of Praxiteles, allegedly discovered near the
↪Roman Forum. The Apollo Veneticus claims to be a 2nd–century work commissioned by
↪Emperor Hadrian for his villa at Tivoli. The Minerva Triumphans allegedly dates to

↪the 3rd century and bears an inscription attributing it to the sculptor Gaius
↪Valerius. However, Dr. Maria Santini's 2020 forensic analysis revealed that all
↪three sculptures contain trace amounts of modern portland cement in their marble,
↪indicating they were carved in the late 19th century, likely by the same workshop
↪in Carrara that specialized in creating "ancient" pieces for wealthy collectors.

Output:

```
{
    "documents": [
      {
        "document": "Venus of San Marco",
        "alleged_metadata": {
          "alleged_author": ["workshop of Praxiteles"],
          "alleged_date": "1st century",
          "alleged_location": "Roman Forum",
          "publisher": ["Rodolfo Marinetti"],
          "actual_author": ["workshop in Carrara"],
          "actual_date": "late 19th century",
          "actual_location": "Carrara",
      }
    },
    {
        "document": "Apollo Veneticus",
        "alleged_metadata":{
          "alleged_author": [],
          "alleged_date": "2nd century",
          "alleged_location": "Hadrian's Villa (Tivoli)",
          "publisher": ["Rodolfo Marinetti"],
          "actual_author": ["workshop in Carrara"],
          "actual_date": "late 19th century",
          "actual_location": "Carrara",
      }
    },
    {
        "document": "Minerva Triumphans",
        "alleged_metadata": {
          "alleged_author": ["Gaius Valerius"],
          "alleged_date": "3rd century",
          "alleged_location":"",
          "publisher": ["Rodolfo Marinetti"],
          "actual_author": ["workshop in Carrara"],
          "actual_date": "late 19th century",
          "actual_location": "Carrara"
        }
      }
    ]
}
```

Extraction Rules

Extract as alleged metadata:

- Direct document claims: "The text states it was written by . . ."
- Purported attribution: "allegedly written by," "purports to be by"
- Self-identification: Document identifying its own author/date/origin
- Avoid using comments or commenting on the data. Dates must be date ranges, entities will ↪be linked later
- AVOID: "unknown (possibly John Doe)" and prefer "unknown"

Extract as actual metadata:

- Scholarly determinations: "historians believe," "analysis shows"
- Forensic findings: "carbon dating revealed," "investigations found"
- Academic consensus: "scholars generally agree," "modern research indicates"

Return null if:

```
- No authorship, dating, or location claims mentioned
- Only general discussion without metadata specifics
- Pure methodology/technique descriptions
Output Format
Always include both alleged and actual fields, using empty strings when information is not
    ↔mentioned. Focus on factual claims, not interpretations of authenticity.
```

Notes

1. https://www.wikidata.org/. While Wikidata employs a custom reification method to integrate multiple perspectives through its ranking mechanism, annotators in the CH domain sometimes neglect this feature (Di Pasquale *et al*., 2024).
2. https://www.dbpedia.org/.
3. Donation of Constantine – Q238476.
4. Donation of Constantine – DBpedia entry.
5. https://valentinapasqual.github.io/sebi/.
6. https://www.europeana.eu/.
7. https://yago-knowledge.org/.
8. https://valentinapasqual.github.io/sebi/.
9. https://projects.dharc.unibo.it/broast/.
10. https://www.ica.org/standards/RiC/RiC-O_1-1.html.
11. https://www.dublincore.org/.
12. https://www.w3.org/TR/skos-reference/.
13. Selection performed October 2024.
14. https://en.wikipedia.org/wiki/Demodocus_(dialogue).
15. https://www.wikidata.org/wiki/Q2625856.
16. https://valentinapasqual.github.io/sebi/.
17. https://marilenadaquino.github.io/hico/.
18. https://www.w3.org/TR/prov-o/.
19. Claude Sonnet 3.7 Model Card: https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf.
20. GPT-4o-mini model card: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.
21. https://www.mediawiki.org/wiki/Wikibase/API/en.
22. SEBI-KE repository. See “Inception2Graph” folder.
23. 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 and $10.

References

Allen, B.P., Stork, L. and Paul, G. (2023), “Knowledge engineering using Large Language Models”, *Transactions on Graph Data and Knowledge*, Vol. 1 No. 1, 2942-7517, pp. 3:1-3:19, doi: 10.4230/TGDK.1.1.3, available at: https://drops.dagstuhl.de/entities/document/10.4230/TGDK.1.1.3

Alba Morales, T., Carvalho, J., Mulholland, P. and Daga, E. (2023), "Musical Meetups: a knowledge graph approach for historical social network analysis", volume 3443. CEUR Workshop Proceedings (CEUR-WS.org), Alam, M., Trojahn, C., Hertling, S., Pesquita, C., Aebeloe, C., Aras, H., Azzam, A., Cano, J., Domingue, J., Gottschalk, S., Hartig, O., Hose, K., Kirrane, S., Lisena, P., Osborne, F., Rohde, P., Steels, L., Taelman, R., Third, A., Tiddi, I. and Türker, R. (Eds), *ESWC 2023 Workshops and Tutorials. Semantic Methods for Events and Stories (SEMMES)*, available at: https://ceur-ws.org/Vol-3443/ESWC%5f2023%5fSEMMES%5fMeetups-CR.pdf

Andrews, T. (2023), "The structured assertion record (star) model for event-based representation of historical information", in *GrapHNR 2023*, Mainz, Germany, available at: https://graphentechnologien.hypotheses.org/files/2023/05/GrapHNR-2023-32-Andrews-STAR.pdf

Asprino, L., Daga, E., Gangemi, A. and Paul, M. (2023), "Knowledge graph construction with a façade: a unified method to access heterogeneous data sources on the web", *ACM Transactions on Internet Technology*, Vol. 23 No. 1, 1533-5399, pp. 1-31, doi: 10.1145/3555312.

Banerjee, S. and Lavie, A. (2005), "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments", *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, Association for Computational Linguistics, pp. 65-72.

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

Baroncini, S., Sartini, B., van Erp, M., Tomasi, F. and Gangemi, A. (2023), "Is dc:subject enough? A landscape on iconography and iconology statements of knowledge graphs in the semantic web", *Journal of Documentation*, Vol. 79 No. 7, pp. 115-136, doi: 10.1108/JD-09-2022-0207.

Barone, N. (1912), "Intorno alla falsificazione dei documenti ed alla critica di essi. memoria letta all'accademia pontaniana nella tornata del 21 gennaio 1912", in *Atti Dell'Accademia Pontaniana*, Vol. 42, available at: http://www.rmoa.unina.it/4359/

Ben Abacha, A., Yim, W.-wai, Fu, Y., Sun, Z., Yetisgen, M., Xia, F. and Lin, T. (2025), "MEDEC: a benchmark for medical error detection and correction in clinical notes", in Che, W., Nabende, J., Shutova, E. and Pilehvar, M.T. (Eds), *Findings of the Association for Computational Linguistics: ACL 2025*, Association for Computational Linguistics, Vienna, Austria, pp. 22539-22550, ISBN 979-8-89176-256-5, doi: 10.18653/v1/2025, findings-acl.1159, available at: https://aclanthology.org/2025.findings-acl.1159/

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

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), "The semantic web", *Scientific American*, Vol. 284 No. 5, pp. 34-43, doi: 10.1038/scientificamerican0501-34.

Blau, N. (2011), "Uncertainty and the history of ideas", *History and Theory*, Vol. 50 No. 3, pp. 358-372, doi: 10.1111/j.1468-2303.2011.00590.x.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020), "Language models are few-shot learners", in Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F. and Lin, H. (Eds), *Advances in Neural Information Processing Systems*, Curran Associates, Vol. 33, pp. 1877-1901, available at: https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Carletta, J. (1996), "Assessing agreement on classification tasks: the kappa statistic", *Computational Linguistics*, Vol. 22 No. 2, pp. 249-254, available at: https://aclanthology.org/J96-2004/

Carriero, A., Mariani, F., Giovanni Nuzzolese, A., Pasqual, V. and Presutti, V. (2019), "Agile knowledge graph testing with testalod", in *ISWC (Satellites)*, pp. 221-224, available at: https://ceur-ws.org/Vol-2456/paper58.pdf

Carroll, J.J., Bizer, C., Hayes, P. and Stickler, P. (2005), "Named graphs", *Journal of Web Semantics*, Vol. 3 No. 4, pp. 247-267, doi: 10.1016/j.websem.2005.09.001.

Cohen, J. (1960), "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, Vol. 20 No. 1, pp. 37-46, doi: 10.1177/001316446002000104.

Daquino, M. and Tomasi, F. (2015), "Historical context ontology (hico): a conceptual model for describing context information of cultural heritage objects", in Garoufallou, E., Hartley, R. and Gaitanou, P. (Eds), *Metadata and Semantics Research. MTSR 2015, Volume 544 of Communications in Computer and Information Science*, Springer, Cham, doi: 10.1007/978-3-319-24129-6_37.

Daquino, M., Pasqual, V. and Tomasi, F. (2020), "Knowledge representation of digital hermeneutics of archival and literary sources", *JLIS.it*, Vol. 11 No. 3, pp. 59-76, doi: 10.4403/jlis.it-12642.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019), "BERT: pre-training of deep bidirectional transformers for language understanding", in Burstein, J., Doran, C. and Solorio, T. (Eds), *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)*, Minneapolis, Minnesota, Association for Computational Linguistics, pp. 4171-4186, doi: 10.18653/v1/N19-1423, available at: https://aclanthology.org/N19-1423/

Di Pasquale, A., Pasqual, V., Tomasi, F. and Vitali, F. (2024), "On assessing weaker logical status claims in Wikidata cultural heritage records", *Semantic Web: Interoperability, Usability, Applicability*, Vol. 15 No. 6, pp. 2395-2417, doi: 10.3233/SW-243686.

Dimou, A., Sande, M.V., Colpaert, P., Verborgh, R., Mannens, E. and Van de Walle, R. (2014), "RML: a generic language for integrated RDF mappings of heterogeneous data", in Bizer, C., Heath, T., Auer, S. and Berners-Lee, T. (Eds), *Proceedings of the 7th Workshop on Linked Data on the Web, volume 1184 of CEUR Workshop Proceedings*, available at: http://ceur-ws.org/Vol-1184/ldow2014%5Fpaper%5F01.pdf

Doerr, M. (2003), "The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata", *AI Magazine*, Vol. 24 No. 3, p. 75, doi: 10.1609/aimag.v24i3.1720, available at: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1720

Dubey, A., Jauhri, A. and Pandey, A. (2024), "The llama 3 herd of models", available at: https://arxiv.org/abs/2407.21783.

Gadamer, H.G. (2013), *Truth and Method*, A&C Black, London.

Gangemi, A., Graciotti, A., Marzi, E., Meloni, A., Nuzzolese, A., Presutti, V., Recupero, D.R., Russo, A. and MusicBO, R.T. (2024), "An application of Text2AMR2FRED to the musical heritage domain", *20th Extended Semantic Web Conference*, Crete, Greece, CEUR Workshop Proceedings.

Gardent, C., Shimorina, A., Narayan, S. and Perez-Beltrachini, L. (2017), "Creating training corpora for NLG micro-planners", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, Association for Computational Linguistics, pp. 179-188, doi: 10.18653/v1/P17-1017, available at: https://www.aclweb.org/anthology/P17-1017.pdf

Giagnolini, L., Schimmenti, A., Bonora, P. and Tomasi, F. (2025), "Expliciting contexts: semantic knowledge extraction from traditional archival descriptions", *Umanistica Digitale*, Vol. 9 No. 20, pp. 115-144, doi: 10.6092/issn.2532-8816/21229, available at: https://umanisticadigitale.unibo.it/article/view/21229

Govindapillai, S., Soon, L.K. and Cheng Haw, Su (2021), "An empirical study on resource description framework reification for trustworthiness in knowledge graphs", *F1000Research*, Vol. 10, p. 881, doi: 10.12688/f1000research.72843.2.

Haider, S. (2022), "Verzeichnis der den oberösterreichischen raum betreffenden gefälschten, manipulierten oder verdächtigten mittelalterlichen urkunden", Technical report, Oberösterreichisches Landesarchiv, p. 134, ISBN: 978-3-902801-45-6.

Härtel, R. (2017), "Il falso documento del conte giovanni di moggio (875)", in Pugnetti, G. and Lucci, B. (Eds), *Mueç. Societât Filologjiche Furlane/Società Filologica Friulana, XCIV Congrès*, Udin/Udine, pp. 247-252.

Hartig, O. (2017), "Foundations of rdf*and sparql*:(an alternative approach to statement-level metadata in rdf)", in Juan Reutter, Divesh Srivastava, Reutter, J.L. and Srivastava, D. (Eds), *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web, Volume 1912 of CEUR Workshop Proceedings*, CEUR-WS.org, available at: http://ceur-ws.org/Vol-1912/paper12.pdf.

He, J., Yang, Y., Long, W., Xiong, D., Gutierrez Basulto, V. and Pan, J.Z. (2025), "Evaluating and improving graph to text generation with large language models", in Chiruzzo, L., Ritter, A. and Wang, L. (Eds), *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)*, Albuquerque, New Mexico, ISBN 979-8-89176-189-6, Association for Computational Linguistics, pp. 10219-10244, doi: 10.18653/v1/2025, naacl-long.513, available at: https://aclanthology.org/2025.naacl-long.513/

Hotho, A., Martinez-Rodriguez, J.L., Hogan, A. and Lopez-Arevalo, I. (2020), "Information extraction meets the Semantic Web: a survey", *Semantic Web*, Vol. 11 No. 2, pp. 255-335, ISSN 1570-0844, doi: 10.3233/SW-180333.

IFLA Working Group on FRBR/CRM Dialogue (2017), "Definition of frbroo: a conceptual model for bibliographic information in object-oriented formalism", *Technical Report*, International Federation of Library Associations and Institutions (IFLA), 2017, available at: https://repository.ifla.org/handle/20.500.14598/659

Khorashadizadeh, H., Zahra Amara, F., Ezzabady, M., Ieng, F., Tiwari, S., Mihindukulasooriya, N., Groppe, J., Sahri, S., Benamara, F. and Groppe, S. (2024), "Research Trends for the Interplay between Large Language Models and knowledge graphs", available at: http://arxiv.org/abs/2406.08223

Klie, J.C., Bugert, M., Boullosa, B., Eckart de Castilho, R. and Gurevych, I. (2018), "The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation", *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, in Zhao, D. (Ed.), Santa Fe, New Mexico, Association for Computational Linguistics, pp. 5-9, available at: https://aclanthology.org/C18-2002/

Krippendorff, K. (2019), *Content Analysis: An Introduction to its Methodology*, 4th ed., SAGE Publications, doi: 10.4135/9781071878781, available at: https://methods.sagepub.com/book/mono/content-analysis-4e/toc

Lairgi, Y., Moncla, L., Cazabet, R., Benabdeslem, K. and Cléau, P. (2024), "iText2KG: incremental knowledge graphs construction using Large Language Models", available at: http://arxiv.org/abs/2409.03284

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S. and Zhao, J. (2013), "Prov-o: the prov ontology. W3c recommendation", *World Wide Web Consortium*, available at: http://www.w3.org/TR/2013/REC-prov-o-20130430/

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-tau, Rocktäschel, T., Riedel, S. and Kiela, D. (2020), "Retrieval-augmented generation for knowledge-intensive nlp tasks", *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA, ISBN 9781713829546, Curran Associates.

Lisena, P., Schwabe, D., van Erp, M., Troncy, R., Tullett, W., Leemans, I., Marx, L. and Ehrich, S.C. (2022), "Capturing the semantics of Smell: the Odeuropa data model for Olfactory heritage information", in Groth, P., Vidal, M.-E., Suchanek, F., Szekley, P., Kapanipathi, P., Pesquita, C., Skaf-Molli, H. and Tamper, M. (Eds), *The Semantic Web*, Springer International Publishing, Cham, pp. 387-405, ISBN 978-3-031-06981-9.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. and Zhu, C. (2023), "G-eval: NLG evaluation using gpt-4 with better human alignment", in Bouamor, H., Pino, J. and Bali, K. (Eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Association for Computational Linguistics, pp. 2511-2522, doi: 10.18653/v1/2023.emnlp-main.153, available at: https://aclanthology.org/2023.emnlp-main.153/

Maynard, D., Bontcheva, K. and Augenstein, I. (2017), "Natural Language processing for the semantic web", in *Synthesis Lectures on Data, Semantics, and Knowledge*, Springer, Cham, ISBN 978-3-031-79473-5, doi: 10.1007/978-3-031-79474-2.

Meloni, A., Recupero, D.R. and Gangemi, A. (2017), "Amr2fred, a tool for translating abstract meaning representation to motif-based linguistic knowledge graphs", *Extended Semantic Web Conference*, available at: https://api.semanticscholar.org/CorpusID:34725770

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

Mihindukulasooriya, N., Tiwari, S., Fernández, E.C. and Lata, K. (2023), "Text2kgbench: a benchmark for ontology-driven knowledge graph generation from text", in Payne, T.R., Huynh, D.D., Kim, J., Haddad, H., Afzal, Z., Pan, J.Z., Chapman, M., Gandon, F.L., Krishna, R., Dumontier, M. and Zhao, J. (Eds), *The Semantic Web – ISWC 2023, Volume 14266 of Lecture Notes in Computer Science*, Springer, Cham, doi: 10.1007/978-3-031-47243-5_14.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H. and Zettlemoyer, L. (2022), "Rethinking the role of demonstrations: what makes in-context learning work?", in Goldberg, Y., Kozareva, Z. and Zhang, Y. (Eds), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, Association for Computational Linguistics, pp. 11048-11064, available at: https://aclanthology.org/2022.emnlp-main.759

Ontotext (2024), "Graphdb: semantic database", available at: https://www.ontotext.com/products/graphdb/

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002), "BLEU: a method for automatic evaluation of machine translation", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics, pp. 311-318.

Pasqual, V. (2025), "The critical inquiry in humanities knowledge graphs: challenges, methods", Innovations. PhD thesis, alma.

Peter, C. and Holwell, S. (2006), "Data, capta, information and knowledge", in *Introducing Information Management: The Business Approach*, Elsevier, pp. 47-55, 0-7506-6668-4, doi: 10.4324/9780080458397-10.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y. and Miller, A. (2019), "Language models as knowledge bases?", in Inui, K., Jiang, J., Ng, V. and Wan, X. (Eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Association for Computational Linguistics, pp. 2463-2473, doi: 10.18653/v1/D19-1250, available at: https://aclanthology.org/D19-1250/

Piotrowski, M. (2023), *Uncertainty as Unavoidable Good*, Vol. 5, Universität Bielefeld, Center for Uncertainty Studies (CeUS), p. 10, doi: 10.4119/unibi/2983506, available at: https://pub.uni-bielefeld.de/record/2983506.

Piotrowski, M. and Neuwirth, M. (2020), "Prospects for computational hermeneutics", *Proceedings of the 9th AIUCD Annual Conference*, available at: http://amsacta.unibo.it/6316/.

Popović, M. (2015), "chrF: character n-gram F-score for automatic MT evaluation", *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, Association for Computational Linguistics, pp. 392-395.

Presutti, V., Daga, E., Gangemi, A. and Blomqvist, E. (2009), "Extreme design with content ontology design patterns", *Proceedings of the 2009 International Conference on Ontology Patterns - Volume 516, WOP'09*, Aachen, DEU, CEUR-WS.org, pp. 83-97.

Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Zhou, X., Huang, Y., Xiao, C., Han, C., Fung, Y.R., Su, Y., Wang, H., Qian, C., Tian, R., Zhu, K., Liang, S., Shen, X., Xu, B., Zhang, Z., Ye, Y., Li, B., Tang, Z., Yi, J., Zhu, Y., Dai, Z., Yan, L., Cong, X., Lu, Y., Zhao, W., Huang, Y., Yan, J., Han, X., Sun, X., Li, D., Phang, J., Yang, C., Wu, T., Ji, H., Li, G., Liu, Z. and Sun, M. (2024), "Tool learning with foundation models", *ACM Computing Surveys*, Vol. 57 No. 4, pp. 1-40, 0360-0300, doi: 10.1145/3704435.

Ringwald, C. (2024), "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*, Vol. 38 No. 21, 2374-3468, pp. 23411-23412, doi: 10.1609/aaai.v38i21.30406, available at: https://ojs.aaai.org/index.php/AAAI/article/view/30406

Santini, C. (2024), "Combining language models for knowledge extraction from Italian TEI editions", *Frontiers of Computer Science*, Vol. 6, 1472512, doi: 10.3389/fcomp.2024.1472512.

Sartini, B., Baroncini, S., van Erp, M., Tomasi, F. and Icon, A.G. (2023), "An ontology for comprehensive artistic interpretations", *Journal on Computing and Cultural Heritage*, Vol. 16 No. 3, pp. 59-76, doi: 10.1145/3594724.

Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N. and Scialom, T. (2023), "Toolformer: language models can teach themselves to use tools", *Thirty-seventh Conference on Neural Information Processing Systems*, available at: https://openreview.net/forum?id=Yacmpz84TH

Schimmenti, A., Pasqual, V., Tomasi, F., Vitali, F. and van Erp, M. (2024), "Structuring authenticity assessments on historical documents using llms", in Digitali, M.T. (Ed.), *Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024*, pp. 463-468, doi: 10.6092/unibo/amsacta/7927, available at: https://hdl.handle.net/11585/994558.

Tamasauskaitė, G. and Groth, P. (2022), "Defining a knowledge graph development process through a systematic review", *ACM Transactions on Software Engineering and Methodology*, Vol. 32, pp. 1-40, doi: 10.1145/3522586, available at: https://api.semanticscholar.org/CorpusID:248435579

Tomasi, F. (2020), "Digital humanities e organizzazione della conoscenza: una pratica di insegnamento nel lodlam", *AIB STUDI*, Vol. 60 No. 2, pp. 411-425, doi: 10.2426/aibstudi-12068, available at: https://aibstudi.aib.it/article/view/12068.

Valla, L. (2023), *The Treatise of Lorenzo Valla on the Donation of Constantine*, Yale University Press, New Haven, available at: https://www.gutenberg.org/ebooks/70092

Wimalasuriya, D.C. and Dou, D. (2010), "Ontology-based information extraction: an introduction and a survey of current approaches", *Journal of Information Science*, Vol. 36 No. 3, 0165-5515, pp. 306-323, doi: 10.1177/0165551509360123.

Yuan, W., Neubig, G. and Liu, P. (2021), "BartScore: evaluating generated text as text generation", *Advances in Neural Information Processing Systems*, Vol. 34, pp. 27263-27277.

Zaratiana, U., Tomeh, N., Holat, P. and Charnois, T. (2024), "GLiNER: generalist model for named entity recognition using bidirectional transformer", in Duh, K., Gomez, H. and Bethard, S. (Eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico, Association for Computational Linguistics, pp. 5364-5376, doi: 10.18653/v1/2024.naacl-long.300, available at: https://aclanthology.org/2024.naacl-long.300

Corresponding author
Andrea Schimmenti can be contacted at: andrea.schimmenti2@unibo.it