xch-MIND: Multi-Agent Interpretive Knowledge Graph Augmentation over Cultural Heritage Linked Data

Michele Stingo[1]

[1]Activa Digital, Via Alessandro Severo, 58, 00145 Roma RM, Italia

Abstract

Institutional Linked Data repositories for cultural heritage, such as the Italian ArCo knowledge base, provide rich factual descriptions but lack interpretive layers connecting entities through spatial, temporal, and typological relationships. We present **xch-MIND (eXtended Cultural Heritage — Multi-agent Interpretive Nexus Discovery)**, a multi-agent system that combines LLM-powered interpretive assertion generation with hybrid confidence scoring and non-invasive ontology layering over existing Linked Data. Four specialised agents analyse entities from ArCo, proposing clusters and relations that are validated through a hybrid score combining LLM self-assessment with deterministic domain-specific checks. All assertions are serialised as RDF triples in a purpose-built `xch:` ontology aligned with PROV-O, with an incremental multi-run strategy for progressive knowledge accumulation. Evaluated on 53 dolmen entities over six runs, xch-MIND discovers 175 unique interpretive assertions at 0.92 average confidence, serialised as 24,788 RDF triples with zero SHACL validation errors. Designed as a decision-support tool, the system provides full provenance and a tiered validation workflow that preserves expert authority over knowledge commitments.

Keywords

Knowledge Graph Augmentation, Large Language Models, Multi-Agent Systems, Cultural Heritage, Linked Data, Ontology Alignment, Provenance

1. Introduction

Cultural heritage institutions worldwide have invested significant effort in publishing structured data as Linked Data [1, 2]. Italy's ArCo (Architettura della Conoscenza — Architecture of Knowledge) [3] represents one of the most comprehensive knowledge bases, modelling over one million cultural properties with detailed descriptive, geographic, and administrative metadata. However, institutional ontologies are intentionally conservative: they capture *factual* information (e.g., location, dating, material) but refrain from encoding *interpretive* relationships (i.e., multi-dimensional clusters, temporal contemporaneity, typological similarities), leaving implicit the latent thematic connections and narrative threads that bind isolated entities together.

This "interpretation gap" limits the utility of knowledge graphs (KGs) for domain experts and the emerging *Web Economy* of cultural heritage [4], where holistic data management is crucial for supporting diverse stakeholders—from scholars to tourists. Archaeologists and curators often need to explore hypothetical connections—such as testing if a group of dolmens shares a specific construction technique or if spatially clustered sites are also temporally coeval. While SPARQL enables sophisticated data retrieval, enriching KGs with *de novo* interpretive assertions (i.e., relationships not explicitly present in existing data) requires both semantic modeling and multi-dimensional validation—tasks demanding LLM reasoning grounded in domain knowledge.

Generating such interpretive layers manually is prohibitively expensive and does not scale. We frame this task as *knowledge graph augmentation*: rather than constructing a KG from unstructured text, xch-MIND generates new interpretive triples grounded in existing Linked Data—a complementary perspective to text-to-KG generation that shares the same core challenges of entity resolution, relation extraction, and output validation. Recent advances in Large Language Models (LLMs) have

ESWC'26: 5th Workshop on LLM-Integrated Knowledge Graph Generation from Text (TEXT2KG), May 10-14, 2026, Dubrovnik, Croatia

michele.stingo@activadigital.it (M. Stingo)

0000-0002-0227-6129 (M. Stingo)

demonstrated remarkable capabilities in understanding domain-specific texts, extracting entities and relations, and performing context-aware reasoning [5, 6]. However, applying LLMs directly to knowledge graph construction poses well-known challenges: hallucination [7], lack of grounding to existing vocabularies [8], and absence of auditable provenance [9, 10]. Recent work on hybrid extraction and validation [11] demonstrates that combining LLM reasoning with knowledge graph integration can substantially reduce hallucination artifacts.

In this paper, we present **xch-MIND (eXtended Cultural Heritage — Multi-agent Interpretive Nexus Discovery)**, a system that addresses these challenges through three key design decisions:

1. **Multi-agent decomposition:** We decompose the knowledge graph generation task into specialised cognitive agents, each responsible for one analytical dimension (geographic, temporal, typological, narrative). An orchestrator coordinates agent activation via a configurable control loop, enabling agents to refine outputs based on accumulated context—a pattern validated by recent agentic reasoning frameworks [12, 13].

2. **Hybrid confidence scoring:** Every assertion generated by an LLM agent undergoes algorithmic validation (e.g., Haversine distance checks for spatial claims, period-overlap verification for temporal claims), drawing on reflection patterns [14] as conceptual inspiration for iterative refinement across runs. The hybrid score combines LLM self-assessed confidence with deterministic validators, following best practices in fact verification [15].

3. **Non-invasive ontology layering:** The system generates triples in a dedicated `xch:` namespace that augments—but never modifies—the source ArCo data, with full PROV-O provenance metadata. The ontology is designed as domain-agnostic, enabling transfer to other heritage domains or Linked Data ecosystems with minimal adaptation.

We evaluate xch-MIND on a corpus of 53 dolmen entities from the ArCo catalogue, spanning heterogeneous metadata quality levels. Over six incremental runs, the system produces 175 unique interpretive assertions at 0.92 average confidence, 1,100+ extracted features, and 24,700+ RDF triples with zero SHACL validation errors. This pilot study validates both the multi-agent methodology and the hybrid confidence mechanism, while establishing a foundation for scaling to larger heritage repositories. Throughout its design, xch-MIND operates as a *decision-support system* for domain experts: every assertion carries confidence scores, provenance metadata, and a validation status, enabling archaeologists and curators to inspect, critique, and selectively accept system outputs—rather than treating machine-generated knowledge as authoritative.

The remainder of this paper is structured as follows: Section 2 surveys related work. Section 3 presents the xch-MIND architecture, including the orchestrator and specialised agents. Section 4 reports experimental results. Sections 5 and 6 discuss findings and conclusions.

2. Related Work

This section surveys work across four dimensions: (i) LLM-powered KG construction targeted at interpretive (not merely factual) relationships, (ii) hallucination mitigation through knowledge graph integration, (iii) multi-agent orchestration with cyclic reasoning, and (iv) ontology layering with provenance tracking in cultural heritage.

The integration of Large Language Models (LLMs) into knowledge graph pipelines has predominantly targeted *factual* KG generation—extracting explicit entity-relation triples from unstructured text, as catalogued in recent surveys [5, 16]. However, these approaches often overlook the synthesis of higher-order *interpretive* relationships (spatial, temporal, and typological) that transcend individual facts. xch-MIND addresses this gap by layering interpretive assertions over existing Linked Data rather than constructing KGs from scratch. This process is further complicated by the inherent risk of hallucination in LLM-powered construction [7]. Recent literature systematically analyses KG-based

mitigation strategies: Agrawal et al. [17] categorise approaches leveraging external KGs to ground LLM outputs, while Wagner et al. [18] review methods integrating KGs into inference pipelines, finding significant improvements in factual accuracy. xch-MIND adopts a complementary strategy: rather than grounding generation in external KGs, it couples LLM self-assessed confidence with deterministic validators specialised per analytical dimension (e.g., Haversine distance, period-overlap checks), filtering assertions before they enter the knowledge graph.

Complex reasoning and task decomposition are increasingly addressed through multi-agent architectures with persistent memory [12, 19, 20]. Recent work on cognitive orchestration [21] demonstrates that dynamic knowledge alignment between agents significantly improves collaborative reasoning, while frameworks like Reflexion [14] introduced verbal self-correction via iterative loops. However, none of these frameworks integrate domain-specific algorithmic validators or produce provenance-rich RDF output. xch-MIND implements a multi-pass orchestrator where agents query accumulated discoveries from shared memory, combining task decomposition with iterative self-improvement and deterministic validation.

In the cultural heritage domain, established ontologies like ArCo [3] encode conservative, well-attested metadata aligned with CIDOC-CRM [22] and EDM [23]. Archaeological knowledge graphs increasingly adopt CIDOC-CRM for excavation documentation [24], while national standards like I.PaC [25] establish semantic graph conventions for Italian heritage. Layering interpretive assertions non-invasively over these ecosystems requires dedicated semantic spaces with principled provenance (W3C PROV [9, 10]). Existing solutions either modify source ontologies directly or lack agent-level attribution. xch-MIND defines a dedicated `xch:` namespace with PROV-O integration, maintaining strict RDF interoperability while recording reasoning traces and per-assertion agent attribution—critical for institutional authority in heritage contexts.

Several related efforts address adjacent challenges. Ferilli and Redavid [26] propose GraphBRAIN, a framework for KG management applied to cultural heritage integrating graph-based reasoning with domain knowledge, which lacks multi-agent LLM orchestration. García-Fernández et al. [27] explore human–LLM collaboration for ontology extension, focusing on the extension process itself rather than on autonomous generation. Gola et al. [28] integrate temporal planning with knowledge representation to generate personalised itineraries, a complementary perspective to xch-MIND's narrative path generation. Mulholland et al. [29] demonstrate the need for interpretive layers by enabling domain experts to manually curate paths through a KG of European pipe organs, validating the "interpretation gap" addressed by xch-MIND. Notably, the theme of *incremental formalisation* from their user studies aligns with our multi-run incremental strategy.

Despite these relevant contributions, a significant gap remains: to our knowledge, no existing system combines multi-agent orchestration, hybrid confidence scoring with dimension-specific validators, and non-invasive ontology layering with full provenance. xch-MIND addresses this compound gap by integrating these dimensions into a unified framework.

3. System Architecture

This section describes the xch-MIND architecture, organised into seven components: system overview, agent orchestration, memory management, specialised agents, prompt design, confidence scoring, and ontology design.

3.1. System Overview

The xch-MIND architecture is designed as a three-layered pipeline that transforms raw archaeological Linked Data into validated interpretive knowledge. Illustrated in Figure 1, the system separates concerns into three distinct stages:

Data Ingestion (Blue Layer). This layer establishes the system's empirical grounding by processing RDF/XML exports from ArCo. Data is parsed via the `lxml` Python library against 17 namespace bindings and normalized into a unified Pydantic model (`DolmenEntity`) with 30 typed fields covering

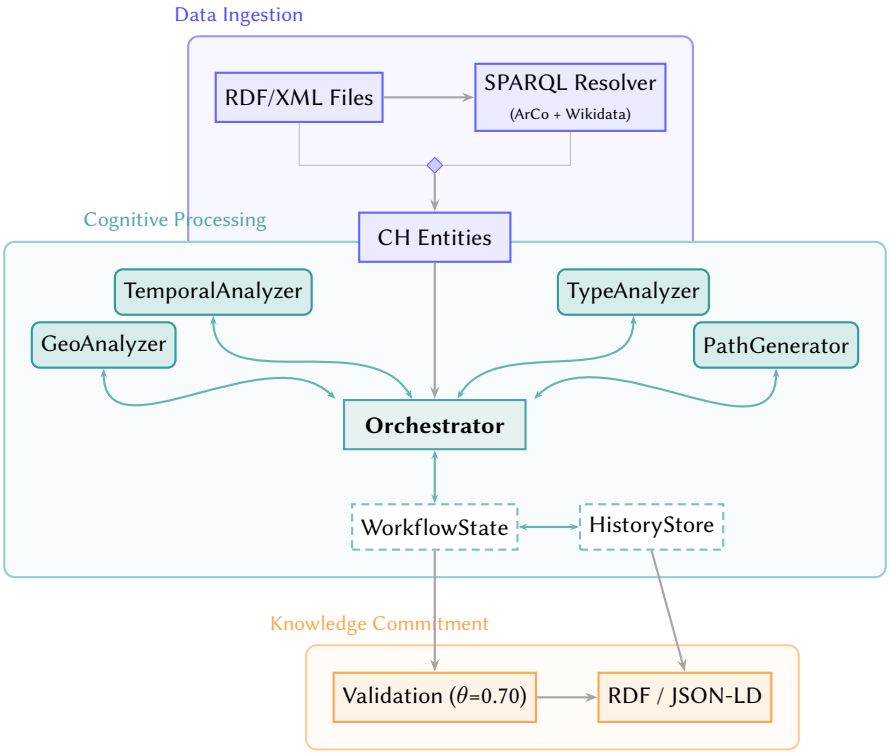

Figure 1: The xch-MIND system architecture, organized into three functional layers: Data Ingestion (blue), multi-agent Cognitive Processing (teal), and Knowledge Commitment (orange).

identification, description, location, and temporality. To complement institutional metadata, the system performs SPARQL resolution for missing coordinates using a dual strategy: querying the ArCo endpoint[1] for GeoSPARQL WKT or `wgs84` properties, with a further fallback to Wikidata via a property path targeting Italian municipalities (*comune of Italy*, `wd:Q747074`). All query results are cached via MD5-keyed JSON storage to ensure reproducibility and performance across runs.

Cognitive Processing (Teal Layer). The core logic resides in a graph-based multi-agent workflow (Section 3.2). Specialized agents query the shared *Session State* for accumulated context and generate interpretive hypotheses through LLM-driven reasoning. Every assertion is subjected to deterministic validation using domain-specific algorithms (e.g., Haversine distance, period overlap) before entering the commitment pipeline. This cyclic interaction allows high-level interpretations—such as thematic paths—to emerge as base-layer assertions are consolidated in memory.

Knowledge Commitment (Orange Layer). The final stage evaluates assertions through hybrid confidence scoring (Section 3.6). Hypotheses meeting the validation criteria are committed to a dedicated `xch:` namespace, layered non-invasively over the source ontology. The resulting output comprises RDF triples with full PROV-O provenance, providing agent-level attribution and reasoning traces—essential for maintaining institutional authority in heritage contexts.

3.2. Agent Orchestration

The orchestrator coordinates agent execution through a graph-based workflow that supports multiple operational modes. The underlying LangGraph topology enables cyclic agent invocation, where each agent pass constitutes a five-phase cognitive cycle.

Upon activation according to the selected strategy, an agent initiates *context retrieval* by querying the Session State (`WorkflowState`) for accumulated entities and pending assertions. This context then primes *hypothesis generation*, where the agent invokes the LLM (Google Gemini 2.5 Flash) with a structured prompt to propose new interpretive triples with self-assessed confidence. To mitigate

[1] https://dati.beniculturali.it/sparql

Algorithm 1: xch-MIND Multi-agent Orchestration: Decomposing knowledge graph generation into specialized agent tasks with hybrid confidence scoring.

Input: entities E, agents A, θ, θ_{review}, strategy, max_runs
Output: Knowledge graph assertions $\mathcal{K}$

$\mathcal{K} \leftarrow \emptyset$; State $\leftarrow$ InitWorkflow(E);
for $p \leftarrow 1$ **to** max_runs **do**
  foreach $a \in$ Orchestrate(A, strategy) **do**
    context $\leftarrow$ RetrieveContext(State, History, a);
    hypotheses $\leftarrow$ LLM.Generate(a.prompt, context);
    foreach $h \in$ *hypotheses* **do**
      $c_{\text{LLM}} \leftarrow h$.confidence;
      $c_{\text{val}} \leftarrow a$.Validate($h$);
      $\Delta \leftarrow |c_{\text{LLM}} - c_{\text{val}}|$;
      $c_{\text{hybrid}} \leftarrow \alpha \cdot c_{\text{LLM}} + \beta \cdot c_{\text{val}} - \max(0,\ (\Delta - 0.4) \cdot \lambda)$;
      if $c_{hybrid} \geq \theta$ **then**
        $\mathcal{K} \leftarrow \mathcal{K} \cup \{h\}$; Commit(History, h, Validated);
      else if $c_{hybrid} \geq \theta_{review}$ **then**
        $\mathcal{K} \leftarrow \mathcal{K} \cup \{h\}$; Commit(History, h, PendingReview);
    end
    UpdateState(State, a, hypotheses);
  end
end
return $\mathcal{K}$;

hallucination, these hypotheses are immediately subjected to *deterministic validation* using domain-specific algorithms (e.g., Haversine distance, period overlap). Finally, in the *commitment phase*, assertions are classified into three tiers—Validated ($\geq \theta$), PendingReview ($\geq \theta_{\text{review}}$), or Rejected—based on their hybrid confidence score (c_{hybrid}), with successful candidates persisted to the *History Store* alongside full reasoning traces (Section 3.6). The orchestration loop executes for a fixed number of runs (*max_runs*), exploiting LLM stochasticity to maximize the coverage of latent interpretive connections across multiple passes.

The architecture supports multiple operational modes through configurable orchestration strategies. The current implementation provides *Comprehensive mode*, which executes all four agents in a deterministic sequence (Geo $\rightarrow$ Temporal $\rightarrow$ Type $\rightarrow$ Path), ensuring systematic coverage of all analytical dimensions. Each agent executes once per pipeline run; the cyclic graph structure enables incremental knowledge accumulation *across* successive runs rather than within a single execution. This design exploits the stochastic nature of LLM generation: repeated pipeline invocations discover connections missed in earlier runs.

The graph topology was deliberately designed to support future operational modes: (i) *Autonomous mode*—fully delegating agent selection and iteration to the multi-agent system, with true intra-run cycles until saturation (implemented but not evaluated in this study); and (ii) *Collaborative mode*—enabling expert-guided analysis through interactive dialogue. Both modes are discussed further in Section 6.

To ensure robustness, each agent includes error-handling mechanisms that manage malformed LLM outputs or service interruptions, ensuring the pipeline's stability and graceful degradation—producing structurally valid but reduced outputs—during analysis.

3.3. Memory Architecture

To support its multi-pass incremental strategy, xch-MIND employs a dual-memory architecture that distinguishes between volatile session-scoped context (Short-Term Memory) and persistent knowledge

(Long-Term Memory).

The **Short-Term Memory (STM)** acts as the system's active reasoning space. Realised via LangGraph's `WorkflowState`, this volatile typed dictionary maintains the current session's entity tracking, pending assertions, agent completion flags, and execution messages. This shared state enables essential cross-agent context sharing within a single execution; for instance, the *PathGenerator* can access geographic clusters identified by earlier agents to ground its narrative sequences. This state is volatile and discarded at the end of each session.

This session-scoped context is formalised into **Long-Term Memory (LTM)** through the `KnowledgeHistoryStore`, which persists validated clusters, relations, and paths into a unified JSON history across pipeline runs. Knowledge accumulation is governed by a dual strategy that combines "hard" post-hoc filtering with "soft" contextual injection. The **NoveltyDetector** implements the hard filter by checking proposed assertions against historical data using Jaccard similarity (defined as the ratio of the intersection to the union of two entity sets) for clusters (threshold > 0.8, chosen to require substantial entity overlap before discarding a proposal as redundant, thereby avoiding false-positive filtering of clusters that share only peripheral members), exact bidirectional matching for relations (checking both $A \to B$ and $B \to A$), and the overlap between the sets of entities (the 'stops') composing each path. In the latter case, the overlap score is increased by a $1.2\times$ multiplier if the thematic labels of the paths are identical, and any path exceeding the 0.7 similarity threshold is then discarded as a duplicate to ensure narrative diversity. Historical assertions below the commitment threshold θ are excluded from novelty detection, enabling a *self-improvement cycle*: since sub-threshold assertions are invisible to the novelty filter, the system can re-propose an improved version in a later run when additional context leads to higher confidence.

The **ContextBuilder** module provides the complement to this process through "soft" context injection. It appends structured markdown sections (headered "EXISTING KNOWLEDGE FROM PREVIOUS RUNS") to LLM task prompts, listing previously discovered knowledge while enforcing display limits (e.g., 10 clusters per type, 20 relations) to maintain token budgets. This proactive injection guides the LLM to add value to existing knowledge, acting as a soft semantic boundary which is then reinforced by the `NoveltyDetector`'s hard post-hoc filtering.

3.4. Specialised Analytic Agents

Four agents decompose the interpretive task along orthogonal dimensions. Each agent encapsulates domain-specific logic for hypothesis generation and validation.

The **GeoAnalyzer** identifies spatial clusters and proximity relations by operationalising the concept of the "archaeological micro-region". It processes WGS84 coordinates from ArCo, computing pairwise Haversine distances to group entities into clusters within a radius $r \leq 10$ km—a threshold reflecting territorial units likely to share functional or cultural ties in antiquity. For broader regional connections, it identifies proximity (`nearTo`) relations using a $d \leq 30$ km boundary, establishing a distinct regional tier ($3\times$ the cluster radius) that captures sites within the same broader cultural landscape without conflating them into a single cluster. Every spatial hypothesis is verified against a deterministic validator to ensure geographic consistency, regardless of the LLM's self-assessed confidence.

The **TemporalAnalyzer** bridges the gap between heterogeneous period labels (e.g., "Età del Bronzo", "Bonnanaro culture") by performing LLM-assisted normalisation into ISO-8601 intervals. Recognising the inherent imprecision of prehistoric chronologies, where boundaries are conventional approximations, the agent identifies contemporaneity (`contemporaryWith`) by allowing a 100-year buffer (τ) during overlap detection. Validated intervals ensure that non-overlapping periods are rejected, while those within the $\pm\tau$ tolerance are materialised as `xch:ChronologicalCluster` or relative temporal assertions.

The **TypeAnalyzer** focuses on the architectural and material nuances hidden in unstructured entity descriptions from ArCo. It extracts features across four categories—architectural, functional, contextual, and material—which are then validated via SpaCy embedding cosine similarity. To ensure cross-run label stability, a dedicated feature normaliser performs syntactic canonicalisation (lowercasing and

English lemmatisation via SpaCy `en_core_web_lg`) followed by semantic deduplication, merging equivalent labels (e.g., "corridors" and "corridor") using a $\theta_{\text{dedup}} = 0.85$ threshold. Clusters are then formed based on a similarity threshold sim > 0.7, tuned to balance precision against semantic variations (e.g., "scavato nella roccia" versus "rock-hewn"). Features appearing in ≥ 2 entities are materialised as `xch:ExtractedFeature` instances.

Acting as the system's narrative synthesiser, the **PathGenerator** weaves accumulated clusters into thematic or evolutionary sequences. It distinguishes between geographic itineraries, chronological progressions, and typological or mixed narrative paths, subclassing them within the `xch:ThematicPath` hierarchy. To ensure grounding, the generator requires at least two base-layer clusters as input. While geographic paths must satisfy a total distance budget of 200 km with consistent leg lengths, chronological and typological paths are validated more flexibly, requiring only verified entity existence to connect thematically related sites across broader regions.

Table 1 summarises agent capabilities.

Table 1
Agent capabilities, validation strategies, and threshold rationale.

Agent	Output	Validation	Rationale
GeoAnalyzer	GeoCluster, nearTo	Haversine $r \leq 10$ km, $d \leq 30$ km	Micro-/regional
TemporalAn.	ChronoCluster, contemporaryWith	Overlap $\pm\tau = 100$ yr	Chrono. imprec.
TypeAnalyzer	TypeCluster, similarTo, Feature	Embed. sim > 0.7	Semantic sim.
PathGenerator	Geo/Chrono/Typo/ NarrativePath	≥ 2 clusters; geo ≤ 200 km	Type-dep.

3.5. Prompt Design

Agent–LLM communication follows a two-tier prompt architecture designed to decouple the agent's analytical persona from its immediate task payload. A stable *system prompt* establishes the domain-specific expert role and reasoning guidelines, while a dynamic *task prompt* injects entity data, runtime constraints, and historical context retrieved from memory.

These system prompts are grounded in archaeological reasoning, adopting chain-of-thought principles [30] to ensure transparent analytical steps. For the *GeoAnalyzer*, the prompt directs the model to prioritize territorial coherence over mere proximity, focusing on sites that would have shared a cultural landscape in antiquity. The *TemporalAnalyzer* is anchored by an embedded six-period chronological framework (spanning from the Neolithic ≈6000 BCE to the Iron Age ≈900 BCE), which acts as a semantic rail to prevent hallucinated dates during interval normalization. To ensure technical interoperability, the *TypeAnalyzer* prompt mandates English-language feature labels, supported by explicit translation patterns (e.g., "*lastrone*→slab"), while the *PathGenerator* utilizes concrete selection rules to prioritize specific narrative subtypes over more generic classifications.

To maintain focus and prevent token-budget overflow, task prompts undergo dynamic constraint injection at runtime. Parameters defined in the system configuration—such as maximum cluster counts, membership bounds, and path length ranges—are interpolated directly into the prompt text as strict operational limits. This guidance is finalized by a structured output requirement: agents request JSON responses conforming to specific Pydantic schemas (e.g., `GeoAnalysisResponse`). This process is reinforced by a robust parsing pipeline that utilizes LangChain's `with_structured_output()` and an automated JSON repair module to handle unbalanced brackets or truncated elements, ensuring that even imperfect model outputs are safely recovered and validated before entering the knowledge graph.

3.6. Hybrid Confidence Scoring

Each assertion undergoes hybrid confidence scoring combining LLM self-assessment with algorithmic validation:

$$c_{\text{hybrid}} = \alpha \cdot c_{\text{LLM}} + \beta \cdot c_{\text{val}} - \max\big(0,\ (\Delta - 0.4) \cdot \lambda\big) \tag{1}$$

where $c_{\text{LLM}} \in [0, 1]$ is the model's self-reported confidence, $c_{\text{val}} \in [0, 1]$ is the validation score (e.g., normalised distance for GeoAnalyzer, overlap ratio for TemporalAnalyzer), $\Delta = |c_{\text{LLM}} - c_{\text{val}}|$, and λ controls the penalty slope for large divergences.

The selection of coefficients ($\alpha = 0.70$, $\beta = 0.30$, $\lambda = 0.30$) reflects a design choice where the LLM's semantic reasoning dominates because interpretive assertions require a level of contextual synthesis that transcends simple algorithmic verification. The validation coefficient provides a corrective signal, downweighting assertions that fail deterministic checks. The penalty term addresses overconfident hallucinations: when $\Delta > 0.4$, a penalty proportional to the excess divergence is applied, so that mildly discrepant assertions receive a small correction while severely discrepant ones are penalised more heavily. This base score is further refined by type-specific modifiers: cluster assertions receive a size-based bonus (rewarding well-populated clusters), while path assertions incur a complexity decay for sequences exceeding five stops. `ExtractedFeature` assertions—which lack an independent LLM confidence signal—receive a prevalence-based score: $c_{\text{feat}} = \min(0.95,\ c_{\text{base}} + \min(0.15,\ p \cdot 0.30) + b_{\text{cat}})$, where $c_{\text{base}} = 0.70$, p is the fraction of entities exhibiting the feature, and $b_{\text{cat}} = 0.05$ for architectural/material categories (zero otherwise). The floor at $c_{\text{base}} = 0.70$ reflects a deliberate design choice: extracted features are grounded in entity descriptions already validated by institutional cataloguers, providing a reliable baseline; the prevalence and category bonuses then modulate this floor upward.

Rather than applying a hard threshold that discards borderline assertions, xch-MIND classifies each assertion into one of three `xch:ValidationStatus` tiers based on c_{hybrid}:

- **Validated** ($c_{\text{hybrid}} \geq 0.70$): committed to Long-Term Memory; high-confidence assertions ready for downstream use.
- **PendingReview** ($0.60 \leq c_{\text{hybrid}} < 0.70$): committed but flagged for expert review, preserving potentially valuable interpretations that fall below the auto-acceptance boundary.
- **Rejected** ($c_{\text{hybrid}} < 0.60$): excluded from the knowledge graph; assertions with insufficient grounding.

This tiered scheme aligns with the system's design as a decision-support tool (Section 5): rather than silently discarding borderline assertions, it surfaces them with appropriate metadata for domain experts to adjudicate.

3.7. Ontology Design: xch Namespace and PROV-O Integration

Rather than modifying ArCo, xch-MIND introduces a dedicated `xch:` namespace[2] that augments institutional data with interpretive assertions through a non-invasive layering pattern. This design preserves source integrity while enabling gradual enrichment. At the heart of this architecture lies the `xch:InterpretiveAssertion`, an abstract superclass (`rdfs:subClassOf prov:Entity`) from which four primary hierarchies derive: (i) **Relations**—`SpatialRelation`, `TemporalRelation`, and `TypologicalRelation` for pairwise entity links; (ii) **Clusters**—`GeographicCluster`, `ChronologicalCluster`, and `TypologicalCluster` for entity groupings; (iii) **Thematic Paths**—`GeographicPath`, `ChronologicalPath`, `TypologicalPath`, and `NarrativePath`, composed of ordered `PathStop` instances; and (iv) **Extracted Features**—`ArchitecturalFeature` and `FunctionalFeature` for entity characteristics.

The actor-based nature of the system is reflected in the `xch:Agent` class (`rdfs:subClassOf prov:Agent`), which features specific specialisations for each analytical dimension. To track the lifecycle of knowledge, a `ValidationStatus` class supports the tiered workflow—Validated, PendingReview, ManuallyValidated, and Rejected—discussed in Section 3.6. As illustrated in Figure 2, every assertion is

[2] https://w3id.org/xch-mind/ontology/

anchored to its source entities via `xch:groundedOn`, while core properties provide essential metadata, including `xch:confidenceScore` (xsd:decimal), `xch:generatedAt/validatedAt` (xsd:dateTime), and `xch:generatedBy/validatedBy` linking to agents. Intrinsic relationships, such as contemporaneity (`xch:contemporaryWith`) or proximity (`xch:nearTo`), are complemented by bidirectional feature links (`xch:hasFeature/xch:featureOf`) that ensure semantic consistency.

To facilitate downstream consumption, relations are serialised in dual form: as direct OWL object properties (e.g., `arco:X xch:nearTo arco:Y`) for rapid SPARQL traversal, and as reified instances (e.g., `xch:SpatialRelation`) with `xch:source/xch:target` properties that allow for individualised confidence scores and reasoning traces. The `contemporaryWith` property is emitted symmetrically, and geographic clusters include a GeoSPARQL centroid node with `wgs84:lat/long` coordinates for spatial indexing. This architectural bridge is finalised through alignment with W3C PROV-O, where properties like `xch:generatedBy` and `xch:derivedFrom` are declared as sub-properties of `prov:wasGeneratedBy` and `prov:wasDerivedFrom` respectively [9]. This ensures that standard provenance tooling can retrieve xch-MIND metadata without namespace-specific logic, as demonstrated in the practical namespace structure shown in Listing 1.

```turtle
@prefix xch: <https://w3id.org/xch-mind/ontology/> .
@prefix xch-res: <https://w3id.org/xch-mind/resource/> .
@prefix data: <https://w3id.org/xch-mind/data/> .
@prefix arco: <https://w3id.org/arco/resource/ArchaeologicalProperty/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

data:geo_cluster_luras
  a xch:GeographicCluster ;
  rdfs:label "Dolmens of Luras" ;
  xch:includesSite arco:1600389451, arco:1600389452 ;
  xch:regionName "Gallura" ;
  xch:confidenceScore 0.97 ;
  xch:generatedBy xch-res:GeoSpatialAgent ;
  xch:generatedAt "2026-02-12T20:30:00Z"^^xsd:dateTime ;
  xch:hasValidationStatus xch:Validated ;
  xch:groundedOn arco:1600389451, arco:1600389452 .

data:feat_corridor_01
  a xch:ArchitecturalFeature, xch:ExtractedFeature ;
  xch:featureLabel "corridor" ;
  xch:featureOf arco:1600389451 ;
  xch:generatedBy xch-res:TypologicalAgent .

arco:1600389451 xch:hasFeature data:feat_corridor_01 .

data:path_sardinia_01
  a xch:GeographicPath, xch:ThematicPath ;
  rdfs:label "Ancient Dolmens of North Sardinia" ;
  xch:hasTheme data:theme_megalithic ;
  xch:pathLength 3 ;
  xch:hasStop data:stop_1, data:stop_2, data:stop_3 ;
  # Path assertions use xch:derivedFrom (rdfs:subPropertyOf
  # prov:wasDerivedFrom) to trace lineage to source clusters;
  # omitted here for brevity.
```

Listing 1: Representative RDF/Turtle serialisation of the xch-MIND knowledge layer, illustrating the alignment of interpretive clusters and extracted features with PROV-O provenance metadata.

The integrity of this ontological layer is enforced through ten SHACL `NodeShapes` (distributed in `xch-shapes.ttl`) that define strict cardinality and structural constraints beyond syntactic well-formedness. For example, `InterpretiveAssertionShape` requires exactly one `xch:confidenceScore`, exactly one `xch:generatedAt`, and at least one `xch:generatedBy`. Cluster shapes require at least two `xch:includesSite` members, while `PathStopShape` and `ExtractedFeatureShape` enforce specific properties for sequencing and entity grounding. This formal baseline enables a five-level `TripleValidator` to operate across the entire pipeline—verifying structural integrity, namespace bindings, type conformance, property validity, and SHACL consistency via `pyshacl`. This multi-level approach ensures that the zero SHACL errors reported in Section 4 represent genuine ontological conformance rather than vacuous validation, reflecting structural invariants across paths, clusters, and relations.

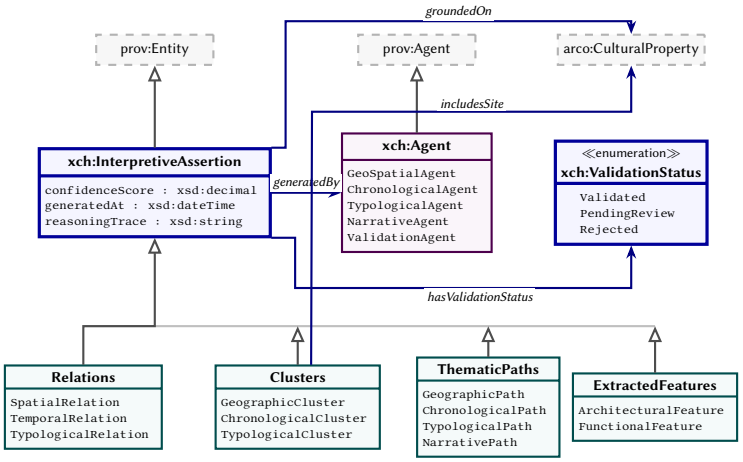

Figure 2: Architectural overview of the `xch:` ontology. The diagram illustrates the hierarchical nesting of interpretive assertions and their strategic anchoring to ArCo and PROV-O standards. Legend: dashed nodes (external classes), open triangles (`rdfs:subClassOf`), solid arrows (`owl:ObjectProperty`).

4. Experimental Evaluation

4.1. Dataset and Technical Configuration

We evaluate xch-MIND on a corpus of 53 dolmen entities from the ArCo catalogue, predominantly distributed across two geographically and culturally distinct regions: Puglia (southern continental Italy) and Sardinia, with one outlier in Liguria. The 53 entities constitute the complete set of dolmen records retrievable from the ArCo `ArchaeologicalProperty` class; dolmens were selected as a typologically homogeneous class with sufficient geographic density for meaningful cluster analysis. This dataset was specifically chosen for its heterogeneity in metadata quality: while 27 entities provide rich textual descriptions detailing architectural features and historical context, the remaining 26 contain only nomenclature and basic coordinates. To address initial data gaps, the Wikidata fallback strategy (Section 3.1) was employed to recover missing WGS84 coordinates for 27 entities that lacked spatial data in the primary catalogue.

The system's cognitive backend is powered by Google Gemini 2.5 Flash, selected for its native structured-output support (reducing JSON parsing failures) and 1M-token context window that accommodates the full entity corpus in a single prompt. The architecture is model-agnostic by design: LLM interactions are mediated through LangChain's abstraction layer, enabling substitution with alternative providers (e.g., OpenAI, Anthropic) without modifying the orchestration logic. The model is configured with a tiered temperature scheduling to balance factual precision with interpretive creativity. Deterministic agents (*GeoAnalyzer*, *TemporalAnalyzer*, and the *Orchestrator*) operate at temperature 0.0 to ensure reproducible factual output, while interpretive latitude is progressively granted to the *TypeAnalyzer* (0.2) for pattern recognition and the *PathGenerator* (0.4) for narrative synthesis. Linguistic processing is supported by a dual SpaCy pipeline: `it_core_news_lg` for computing cosine similarity over Italian source descriptions, and `en_core_web_lg` for the lemmatisation and deduplication of English-language feature labels. The pipeline was executed in six cumulative runs over a 48-hour period, with a hybrid confidence threshold set at $\theta = 0.70$.

4.2. Cumulative Knowledge Discovery

The experimental results, summarized in Table 2, demonstrate the system's capacity to formalise complex interpretive knowledge with high reliability ($c_{\text{avg}} = 0.92$ for interpretive assertions).

Across 1,497 committed assertions, xch-MIND successfully identified 14 geographic micro-regions and 15 chronological clusters. A key success is the identification of the "Dolmens of Luras" cluster in Sardinia (grouping sites Arzoledda, Bilella, Ciuledda, and Ladas within a 2 km radius) and the

Table 2
Cumulative pipeline output (6 runs, 53 entities).

Assertion Type	Count	Avg. Confidence
Geographic Cluster	14	0.97
Proximity Relation (`nearTo`)	25	0.96
Chronological Cluster	15	0.93
Temporal Relation (`contemporaryWith`)	15	0.95
Typological Cluster	29	0.86
Similarity Relation (`similarTo`)	52	0.90
Thematic Path	25	0.93
Interpretive Assertions	**175**	**0.92**
Extracted Feature	1,171	0.82
Path Stop	151	—
Total Committed Assertions	**1,497**	—
Total RDF Triples	**24,788**	—
Validation Errors	**0**	—

Table 3
Per-run assertion generation showing incremental discovery and novelty filtering.

Run	Interp.	Features	Triples	Primary Contributions
1	82	192	5,198	Geo (29), chrono (21), typo (27), paths (5)
2	28	179	3,859	Geo (10), typo (13), paths (5)
3	19	179	3,566	Typo (15), paths (4)
4	27	195	4,067	Chrono (9), typo (14), paths (4)
5	11	200	3,864	Typo (8), paths (3)
6	8	226	4,234	Typo (4), paths (4)
Total	**175**	**1,171**	**24,788**	

"Dolmens of Bisceglie" cluster in Puglia (comprising Chianca, Frisari, and Albarosa), both of which were assigned high hybrid confidence scores (1.0 and 0.96, respectively), confirmed by deterministic Haversine validation against ArCo coordinates.

The system's capacity for cross-regional temporal reasoning is further evidenced by the successful mapping of Puglia's "Early Bronze Age" to Sardinia's "Bonnanaro culture" (spanning 2300–1700 BCE), creating 15 valid temporal links with $c = 0.95$. Simultaneously, the *TypeAnalyzer* elicited 1,171 fine-grained features across architectural, functional, and material categories, grouping them into 29 typological clusters. These findings were ultimately synthesised into 25 thematic paths, such as the "Megalithic Evolution of Sardinia," which demonstrate the system's ability to generate higher-order interpretive structures from base-layer assertions. The absence of validation errors across all executions confirms the structural robustness of the hybrid scoring and SHACL enforcement layers.

4.3. Run-by-Run Analysis

Table 3 details the incremental knowledge accumulation across six pipeline executions. The observed decline in new interpretive assertions ($82 \rightarrow 8$) reflects the progressive saturation of *structured* dimensions (geographic and chronological), which reach full discovery by the third run. Typological features, by contrast, continue to emerge throughout the experiment, benefiting from the richer combinatorial space of architectural descriptions—consistent with their higher novelty rates (Section 4.5). This growth is managed through a dual novelty-control architecture designed to prevent redundancy:

1. **Contextual Guidance:** Before each LLM invocation, the `ContextBuilder` injects historical

knowledge into the prompt, instructing the model to focus on unexplored patterns. This "soft" guidance steers the model away from already-discovered clusters.

2. **Deterministic Filtering:** Following generation, a post-hoc `NoveltyDetector` computes Jaccard similarity ($J > 0.8$) between proposed and historical assertions. This filter discards redundant clusters regardless of their confidence score.

This proactive orchestration ensures that every committed assertion adds genuine value to the graph. Across the experiment, the `NoveltyDetector` filtered 309 duplicate proposals (17% of the 1,806 total candidates), resulting in the final set of 1,497 unique assertions. Per-agent analysis reveals that filtering is concentrated on finite-dimensional agents (geo: 78.7%, temporal: 75.8%), while the type analyzer (5.1%) and path generator (2.2%) exhibit low filter rates consistent with their richer combinatorial spaces. This asymmetry provides indirect evidence that the Jaccard threshold (> 0.8) avoids over-aggressive filtering: if the threshold were too low, typological clusters—which frequently share peripheral members—would exhibit a substantially higher rejection rate. This behavior validates the multi-run strategy as a principled method for exploiting LLM stochasticity while maintaining strict control over knowledge redundancy.

4.4. Qualitative Analysis

To further illustrate the system's interpretive depth, we examine representative examples across the temporal, typological, and narrative dimensions, where the cognitive synergy between LLM reasoning and algorithmic validation is most evident.

The *TemporalAnalyzer* demonstrated a sophisticated ability to normalise heterogeneous chronological labels. For Sardinian dolmens, terms such as "Età del Rame," "Cultura di Ozieri," and "Neolitico finale" were mapped to their underlying overlapping intervals (approximately 3500–2300 BCE). This enabled the creation of five "Late Neolithic/Eneolithic" clusters with a confidence of 0.96, showcasing how the system can bridge local terminological variations to produce a unified chronological view. Most notably, the agent identified the partial overlap between the Sardinian "Bonnanaro culture" (normalised to 1800–1500 BCE) and the Apulian "Early Bronze Age" (2300–1700 BCE), a cross-regional reasoning task that would typically require significant manual cross-referencing.

In the typological dimension, the *TypeAnalyzer* successfully detected semantic equivalences despite significant lexical variation in the source descriptions. A notable example is the "Rock-Cut Elements" cluster ($c = 0.79$), which linked entities described variously as "scavato nella roccia" (excavated in rock), "struttura ipogeica" (hypogeic structure), and "rock-hewn chamber." By leveraging embedding-based similarity (cosine > 0.7), the system performed cross-lingual matching that identified these features as architecturally equivalent. Similarly, the "Orthostatic Construction" cluster ($c = 0.85$) grouped dolmens characterized by vertical stone slabs, demonstrating the model's ability to extract structural patterns directly from qualitative text.

Finally, the *PathGenerator* synthesised these disparate findings into higher-order narrative sequences. The "Megalithic Evolution of Sardinia" path (10 stops, $c = 0.81$) is a prime example of this synthesis, ordering sites by increasing architectural complexity—from simple orthostatic chambers to elaborate corridor tombs. Likewise, the "Salento Dolmen Trail" (8 stops, $c = 0.76$) proposes a geographic itinerary that connects major site concentrations with isolated monuments in the Murge hinterland. These paths represent more than just spatial routes; they are interpretive constructs that transform traditional database records into curated historical narratives.

4.5. Quantitative Comparison and Sensitivity Analysis

To assess the contribution of individual architectural components, we conduct two complementary experiments: a *single-prompt baseline* and a *sensitivity analysis* on the hybrid confidence mechanism.

Table 4
Multi-agent system vs. single-prompt baseline (single run, same LLM).

Metric	xch-MIND	Baseline
Clusters	29	6
Pairwise Relations	48	4
Thematic Paths (stops)	5 (24)	2 (9)
Extracted Features	192	0
Total Assertions	298	21
LLM Calls	4	1
Duration (s)	238	5

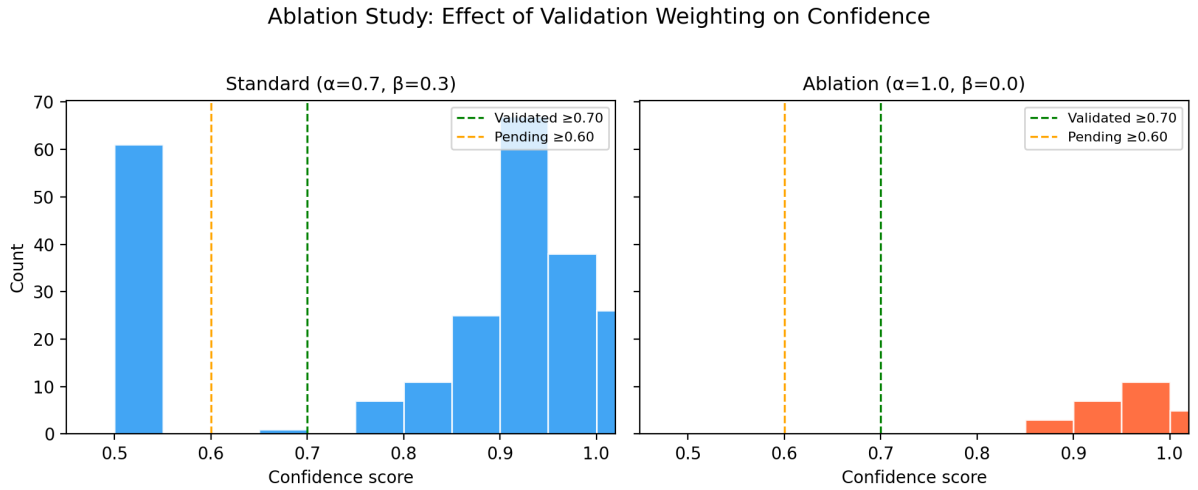

Figure 3: Confidence score distributions under standard configuration ($\alpha = 0.7, \beta = 0.3$) and ablation ($\alpha = 1.0, \beta = 0.0$). Without validation, all scores cluster above 0.85 with no rejected proposals.

Single-Prompt Baseline. We compare xch-MIND against a naïve baseline that sends all 53 entities to the LLM in a single prompt, requesting geographic, chronological, and typological clusters, pairwise relations, and thematic paths simultaneously—without multi-agent decomposition, validation, or memory. Table 4 summarises the results.

The multi-agent architecture produces $14\times$ more assertions than the monolithic baseline. However, this quantitative gap is secondary to the *qualitative* difference: the baseline entirely fails to extract fine-grained features (0 vs. 192), which constitute 64% of xch-MIND's single-run output. This failure is not attributable to fewer LLM calls—even with equivalent compute budget, a single generic prompt lacks the domain-specific output schemas (e.g., `GeoAnalysisResponse`, `TemporalAnalysisResponse`) and dimension-specific validators that enable each agent to elicit structured, verifiable outputs. The specialisation of each agent enables a depth of analysis that is structurally inaccessible to prompt-level task composition. While the baseline is faster (5 s vs. 238 s), the marginal cost of additional LLM calls is negligible relative to this qualitative improvement.

Sensitivity Analysis: Validation Weighting. To isolate the effect of deterministic validation on confidence scoring, we executed a run with validation weight set to zero ($\alpha = 1.0, \beta = 0.0$), removing the algorithmic component from the hybrid formula (Eq. 1). We target the validation component as the primary ablation variable because it represents the system's core discriminative mechanism; memory effects are analysed longitudinally below via the run-by-run data (Table 3), where Run 1 corresponds to a no-memory baseline. Figure 3 shows the resulting confidence distributions.

Under standard configuration, the validation component produces a bimodal distribution: 73.7% of knowledge items are validated ($c \geq 0.70$), while 25.8% are rejected ($c < 0.60$), demonstrating effective

quality discrimination. When validation is removed, all 26 knowledge items receive uniformly high scores ($\bar{c} = 0.937$, $\sigma = 0.045$, min$= 0.85$), with 100% classified as validated and zero rejections. This loss of discriminative power confirms that the deterministic validation component is essential for preventing over-confident outputs—a known failure mode of LLM self-assessment [31].

Ablation: Memory and Novelty Detection. The contribution of the memory architecture can be assessed directly from the run-by-run data in Table 3. Run 1 operates without accumulated knowledge (equivalent to a no-memory condition), while subsequent runs benefit from the `NoveltyDetector`. Per-agent analysis reveals strongly asymmetric filtering: geographic proposals are filtered at 93.5% in runs 2–6 (144/154), temporal proposals at 91.3% (94/103), while typological (6.1%) and path (2.6%) proposals maintain high novelty rates. This pattern confirms that the memory component is essential for preventing redundancy in structured dimensions while preserving the system's capacity for continued discovery in combinatorial spaces.

5. Discussion

Incremental Knowledge Accumulation. The run-by-run analysis presented in Section 4 confirms the effectiveness of incremental learning within the xch-MIND framework. The `NoveltyDetector`'s capacity to filter 17% of total proposals as duplicates demonstrates that the system does not merely accumulate redundant data, but actively manages its cognitive growth. A key result is the asymmetric saturation observed across analytical dimensions: while geographic and chronological patterns are finite and reach full discovery within three runs, typological features maintain a high novelty rate ($\geq$93% per run). This phenomenon is driven by the richer combinatorial space of architectural descriptions and validates our multi-run strategy as a principled approach for exploiting LLM stochasticity for progressive knowledge discovery.

Empowering Domain Experts. xch-MIND is fundamentally designed as a decision-support tool for archaeologists, curators, and heritage scholars, aiming to enhance rather than replace expert judgment. This philosophy marks a significant evolution from traditional manual curation approaches, such as the expert-driven path creation proposed by Mulholland et al. [29]. Where their system requires specialists to manually select and connect cultural objects, xch-MIND automates the generation of candidate interpretive assertions, presenting them to experts for validation through confidence scores and the review queue. For example, verifying a complex grouping like the "Dolmens of Luras" cluster through traditional methods would require an expert to manually query ArCo, compute pairwise distances, and cross-check period labels—a process typically requiring hours. xch-MIND completes this task in seconds while providing full provenance for every link. By transforming the knowledge graph into a "glass box," the system allows experts to judge interpretations rather than perform tedious data cleaning. The `PendingReview` validation state is central to this authority, explicitly flagging low-confidence assertions to ensure that specialists maintain final control over knowledge commitments.

Review Queue and Reproducibility. The operationalisation of this human-in-the-loop workflow is handled by a dedicated post-processing phase at the end of each pipeline run. A SPARQL `SELECT` query identifies all resources carrying the `xch:hasValidationStatus xch:PendingReview` property, extracting relevant metadata such as labels, confidence scores, and reasoning traces into a structured `review_queue.json` file. Sorted by ascending confidence, this file enables experts to adjudicate borderline cases within a streamlined workflow. Once confirmed, an assertion can be promoted to `xch:ManuallyValidated`, a state that records the human decision as a distinct provenance event. Furthermore, transparency is ensured by the production of a `metadata.json` file recording model details, provider information, and category-wise statistics. Together with a *dry-run* mode that simulates the orchestration loop without LLM calls, these artefacts ensure the end-to-end reproducibility of our experimental results.

Confidence Calibration. The sensitivity analysis in Section 4.5 provides empirical grounding for the hybrid confidence weights ($\alpha = 0.7, \beta = 0.3$). When validation is removed, confidence scores collapse

into a narrow high range ($\bar{c} = 0.937, \sigma = 0.045$), eliminating the system's ability to discriminate quality. Under the standard configuration, the distribution is bimodal: relations grounded in deterministic signals (e.g., `nearTo` validated by Haversine distance) achieve $\bar{c} = 0.96$, while purely interpretive assertions (e.g., `similarTo`) exhibit higher variance ($\bar{c} = 0.75, \sigma = 0.20$). This disparity is not a calibration flaw but a feature: it faithfully reflects the epistemic distinction between verifiable and interpretive claims. The penalty term $\max(0, (\Delta - 0.4) \cdot \lambda)$ activates only when LLM and algorithmic assessments diverge significantly ($\Delta > 0.4$), preventing overconfident assertions in the absence of algorithmic support. The 0.3 weight for β was selected to provide sufficient discriminative power while avoiding excessive penalisation of assertions where no deterministic validator applies.

Prompt Adaptability and Domain Transfer. Each agent operates with a parametric prompt comprising a domain-specific role definition (e.g., "You are an expert archaeologist specialising in megalithic monuments") and structured task instructions injected at runtime. Adapting xch-MIND to a new domain involves three progressive levels: (1) *prompt-only* adaptation for domains with analogous spatial and temporal structures (e.g., medieval castles), requiring only role and terminology updates; (2) *prompt and threshold* adaptation for domains with different scales, such as adjusting distance thresholds for urban heritage; (3) *prompt and agent* adaptation for domains requiring new analytical dimensions, such as adding a material composition analyser for museum collections. The orchestration logic, memory architecture, and confidence scoring remain invariant across all three levels.

Conflict Resolution. In the current *comprehensive* mode, inter-agent conflicts are structurally avoided because agents operate sequentially on orthogonal analytical dimensions (spatial, temporal, typological). Each dimension generates independent assertions that coexist in the graph without contradiction—a geographic cluster and a typological cluster may share members, representing complementary rather than conflicting interpretations. The post-hoc `NoveltyDetector` prevents *intra-dimensional* redundancy through Jaccard filtering. For the already-implemented *autonomous* mode, where agents may be invoked in arbitrary order, conflicts are resolved through the confidence-ranked review queue: when overlapping assertions emerge, the expert adjudicates based on the transparency afforded by provenance traces and reasoning metadata.

Limitations and Mitigation Strategies. Despite its performance, several limitations warrant discussion. A broader methodological consideration concerns the absence of formal expert evaluation. We deliberately refrain from reporting precision and recall metrics because no gold-standard dataset exists for interpretive assertions over megalithic structures: unlike factual KG extraction, where source texts provide ground truth, the relationships discovered by xch-MIND—spatial clusters, temporal alignments, narrative paths—are novel hypotheses rather than verifiable extractions. Assertions grounded in deterministic validators—geographic clusters validated by Haversine distance, temporal overlaps by interval arithmetic—are *correct by construction*: their truth value is not a matter of interpretation but of geometric and arithmetic fact (e.g., the "Dolmens of Luras" cluster, whose members lie within 2 km according to ArCo coordinates). These deterministically grounded assertions account for 39% of interpretive outputs (69 of 175). Higher-order interpretive assertions (typological clusters, narrative paths), however, require domain expert adjudication—which is precisely the role served by the tiered validation workflow and review queue (Section 3.6). Formalising this through structured expert evaluation with precision@k metrics constitutes a primary objective for future work.

Beyond this overarching consideration, three specific limitations merit attention. First, extracted features—comprising 78% of committed assertions (1,171 of 1,497)—receive prevalence-based confidence scores (avg. 0.82) that are lower than interpretive assertions (avg. 0.92). To bridge this gap, we envision the introduction of a "Feature Critic" agent to validate architectural vocabularies against external standards like the Getty AAT. Second, we address the calibration of confidence scores. The high average scores for geographic (0.97) and temporal (0.95) relations derive from the design of the hybrid formula, where the penalty term activates only when the disagreement Δ between LLM and algorithm exceeds 0.4. When both agree, scores remain high, reflecting grounded knowledge. However, typological assertions yield lower scores (0.86) due to their interpretive latitude. Future work will involve calibrating these scores against expert assessments through precision@k studies. Third, we acknowledge that system performance correlates with input metadata quality. While entities with rich textual descriptions yield

more complex paths, the system degrades gracefully on sparse data, where geographic and temporal analysis remains effective. Mitigating this through data augmentation via alignment with external knowledge bases like Wikidata or PeriodO represents a key future direction.

Generalisability to Other Domains. The `xch:` ontology was architected as a domain-agnostic layer, focusing on generic interpretive patterns like clustering and path synthesis that are applicable well beyond cultural heritage. Core classes like `InterpretiveAssertion` and `ThematicPath` encode structural relationships that remain valid across various fields. Modifying the system for new domains, such as museum collections or historical archives, requires only targeted updates to agent prompts and input data connectors, while the orchestration logic, memory architecture, and confidence scoring remain entirely reusable. This modularity facilitates rapid domain transfer, allowing xch-MIND to serve as a versatile engine for the semantic augmentation of diverse knowledge bases. We acknowledge that the current evaluation on 53 dolmen entities constitutes a proof-of-concept; validating scalability and cross-domain portability through larger and heterogeneous corpora remains a primary objective for future work.

6. Conclusion and Future Work

We presented xch-MIND, a multi-agent system for the interpretive enrichment of Cultural Heritage Linked Data. By combining the generative power of LLMs with rigorous algorithmic validation and a persistent cognitive cycle, we enable the automated discovery of high-quality spatial, temporal, and typological relationships. Evaluated on 53 dolmen entities from the ArCo knowledge base across six incremental runs, xch-MIND produced 24,788 RDF triples and 1,497 committed assertions with zero SHACL validation errors, demonstrating that LLM-driven augmentation can meet Semantic Web quality standards. The resulting knowledge graph offers a richer, more interconnected view of heritage assets while preserving full provenance and institutional data integrity.

The research roadmap for xch-MIND targets several strategic directions. From an operational perspective, we aim to evaluate the system's already-implemented *Autonomous mode* to measure discovery throughput in dynamically triggered agent cycles. Furthermore, we plan to introduce a *Collaborative mode* to leverage real-time expert dialogue for immediate hypothesis validation. In terms of architectural robustness, we will perform a sensitivity analysis across hybrid confidence thresholds (θ) to identify optimal operating points for varying use cases, while also refining cluster composition to integrate heterogeneous entity types beyond megalithic monuments. Finally, we seek to validate the system's portability by applying the domain-agnostic core of xch-MIND to diverse cultural heritage domains, such as museum collections and historical archives, testing the scalability of our interpretive patterns across the broader digital heritage landscape.

Availability of Material

The xch-MIND codebase, including the ontology definition, agent implementations, and evaluation scripts, is available under an open-source license at https://github.com/6MicheleStingo9/xch-MIND. A persistent snapshot of the software, generated knowledge graph (RDF triples), and evaluation datasets is archived on Zenodo at https://doi.org/10.5281/zenodo.19483661.

Declaration on Generative AI

During the preparation of this work, the author used GitHub Copilot in order to: Grammar and spelling check. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] E. Hyvönen, Publishing and Using Cultural Heritage Linked Data on the Semantic Web, volume 2 of *Synthesis Lectures on the Semantic Web: Theory and Technology*, Morgan & Claypool Publishers, 2012. doi:10.2200/S00452ED1V01Y201210WBE003.

[2] A. Bikakis, E. Hyvönen, S. Jean, B. Markhoff, A. Mosca, Editorial: Special issue on semantic web for cultural heritage, Semantic Web 12 (2021) 163–167. doi:10.3233/SW-210425.

[3] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, C. Veninata, ArCo: The Italian cultural heritage knowledge graph, in: Proceedings of the 18th International Semantic Web Conference (ISWC 2019), Springer, 2019, pp. 36–52. doi:10.1007/978-3-030-30796-7_3.

[4] S. Ferilli, E. Bernasconi, D. Redavid, G. M. Di Nunzio, Knowledge graphs for the web economy: The CHiPS&BITS project on cultural heritage, Umanistica Digitale 10 (2026) 113–138. URL: https://umanisticadigitale.unibo.it/article/view/22159. doi:10.60923/issn.2532-8816/22159.

[5] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599. doi:10.1109/TKDE.2024.3352100.

[6] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Computing Surveys 54 (2021) 71:1–71:37. doi:10.1145/3447772.

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. doi:10.1145/3571730.

[8] A. Amayuelas, J. Sain, S. Kaur, C. Smiley, Grounding LLM reasoning with knowledge graphs, 2025. arXiv:2502.13247.

[9] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, PROV-O: The PROV Ontology, W3C Recommendation, W3C (World Wide Web Consortium), 2013. URL: https://www.w3.org/TR/prov-o/.

[10] L. Moreau, P. Missier, PROV-DM: The PROV Data Model, W3C Recommendation, W3C (World Wide Web Consortium), 2013. URL: https://www.w3.org/TR/prov-dm/.

[11] A. De Santis, M. Balduini, F. De Santis, A. Proia, A. Leo, M. Brambilla, E. Della Valle, Integrating large language models and knowledge graphs for extraction and validation of textual test data, in: The Semantic Web – ISWC 2024: 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11–15, 2024, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2024, pp. 304–323. URL: https://doi.org/10.1007/978-3-031-77847-6_17. doi:10.1007/978-3-031-77847-6_17.

[12] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, Y. Cao, ReAct: Synergizing reasoning and acting in language models, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=WE_vluYUL-X.

[13] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST '23, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3586183.3606763.

[14] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: Language agents with verbal reinforcement learning, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[15] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074/. doi:10.18653/v1/N18-1074.

[16] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, Llms for knowledge

graph construction and reasoning: Recent capabilities and future opportunities, World Wide Web 27 (2024). doi:10.1007/s11280-024-01297-w.

[17] G. Agrawal, T. Kumarage, Z. Alghamdi, H. Liu, Can knowledge graphs reduce hallucinations in LLMs? : A survey, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 3947–3960. URL: https://aclanthology.org/2024.naacl-long.219/. doi:10.18653/v1/2024.naacl-long.219.

[18] R. Wagner, E. Kitzelmann, I. Boersch, Mitigating hallucination by integrating knowledge graphs into LLM inference – a systematic literature review, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 795–805. URL: https://aclanthology.org/2025.acl-srw.53/. doi:10.18653/v1/2025.acl-srw.53.

[19] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan, Tree of Thoughts: Deliberate problem solving with large language models, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[20] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, C. Wang, AutoGen: Enabling next-gen LLM applications via multi-agent conversations, in: First Conference on Language Modeling, 2024. URL: https://openreview.net/forum?id=BAakY1hNKS.

[21] J. Zhang, Y. Fan, K. Cai, X. Sun, K. Wang, OSC: Cognitive orchestration through dynamic knowledge alignment in multi-agent LLM collaboration, in: Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, 2025, pp. 6320–6337. doi:10.18653/v1/2025.findings-emnlp.335.

[22] M. Doerr, The CIDOC conceptual reference model: An ontological approach to semantic interoperability of metadata, AI Magazine 24 (2003) 75–92. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1720. doi:10.1609/aimag.v24i3.1720.

[23] A. Isaac, R. Clayphan, B. Haslhofer, J. Schirrwagen, R. Simon, Europeana Data Model (EDM) Primer, Technical Report, Europeana, 2013. Technical documentation.

[24] G. Hiebel, E. Aspöck, K. Kopetzky, Ontological modeling for excavation documentation and virtual reconstruction of an ancient Egyptian site, ACM Journal on Computing and Cultural Heritage 14 (2021) 32:1–32:14. URL: https://doi.org/10.1145/3439735. doi:10.1145/3439735.

[25] M. Porena, M. Bartoli, L. Cerullo, A. Negri, I.PaC and semantic graphs to represent Italian Cultural Heritage, in: S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, F. Remondino (Eds.), Digital Heritage, The Eurographics Association, 2025. doi:10.2312/dh.20253317.

[26] S. Ferilli, E. Bernasconi, D. Di Pierro, D. Redavid, The GraphBRAIN framework for knowledge graph management and its applications to cultural heritage, in: Bridging the Gap Between AI and Reality: First International Conference, AISoLA 2023, Crete, Greece, October 23–28, 2023, Selected Papers, Springer-Verlag, Berlin, Heidelberg, 2023, pp. 144–161. doi:10.1007/978-3-031-73741-1_10.

[27] J. García-Fernández, J. Verhoosel, J. Ubacht, R. M. Bakker, Ontology engineering with large language models: Unveiling the potential of human-LLM collaboration in the ontology extension process, in: Proceedings of the 4th Workshop on LLM-Integrated Knowledge Graph Generation from Text (TEXT2KG 2025), volume 4020 of *CEUR Workshop Proceedings*, CEUR-WS, 2025, pp. 135–159.

[28] S. Gola, D. Capaldi, A. Chirivì, M. A. Jaziri, L. Leopardi, S. G. Malatesta, I. Muci, A. Orlandini, A. Umbrico, A. Bucciero, Integrating temporal planning and knowledge representation to generate personalized touristic itineraries, in: AIxIA 2024 – Advances in Artificial Intelligence: XXIIIrd International Conference of the Italian Association for Artificial Intelligence, AIxIA 2024, Bolzano, Italy, November 25–28, 2024, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2024, pp. 188–199. doi:10.1007/978-3-031-80607-0_15.

[29] P. Mulholland, P. Van Kranenburg, J. Carvalho, E. Daga, Supporting the end-user curation of cultural heritage knowledge graphs, in: Proceedings of the 35th ACM Conference on Hypertext

and Social Media, HT '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 35–44. URL: https://doi.org/10.1145/3648188.3675132. doi:10.1145/3648188.3675132.

[30] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[31] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. B. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, J. Kaplan, Language models (mostly) know what they know, 2022. arXiv:2207.05221.