STATI GENERALI DEL PATRIMONIO INDUSTRIALE 2026

a cura di
Edoardo Currà, Fabio Fatiguso, Antonella Guida, Antonio Monte
Marina Docci, Graziella Bernardo, Elena Cantatore, Alessandro Mattioli, Claudio Menichelli

GANGEMI EDITORE®
SPA
INTERNATIONAL

Gangemi Editore spa
Via Giulia 142, Roma
www.gangemieditore.it

Le nostre edizioni sono disponibili in Italia e all'estero anche in versione ebook.
Our publications, both as books and ebooks, are available in Italy and abroad.

ISBN 978-88-492-5487-7

DOI: https://cdn.gangemieditore.com/DOI/10.61020/9788849254877.pdf

In copertina: © Mariano De Angelis, Ex Tabacchificio Fortunato Farina, Battipaglia (SA), 2019.

HISTORICAL DOCUMENTS TO SEMANTIC KNOWLEDGE MODELS: AN AI WORKFLOW FOR INDUSTRIAL HERITAGE

Da documenti storici a modelli di conoscenza semantica: un flusso di lavoro basato sull'intelligenza artificiale per il patrimonio industriale

Cassia De Lian Cui[1], Stefano Cursi[2], Davide Simeone[3], Antonio Fioravanti[1], Edoardo Currà[1]

1: Sapienza Università di Roma
2: Istituto di Scienze del Patrimonio Culturale – CNR
3: Università di Brescia

Keyword
AI Assistant; Ontology-Based Systems; Data Integration; Industrial Heritage Documentation; Interoperability.
Assistente AI; sistemi basati sull'ontologia; integrazione dei dati; documentazione del patrimonio industriale; interoperabilità.

Abstract *The ongoing digital revolution has profoundly impacted industry and society driving the urgency to reconsider and innovate current industrial heritage recovery and valorization activities. Indeed, the industrial heritage field faces increasing challenges related to the management and interpretation of historical data, much of which is unstructured, dispersed, and difficult to integrate into modern conservation practices, requiring high expertise and manual work for structuring unorganized information into digital knowledge bases and information models. This research explores how the tangible and immaterial information can be processed and integrated into an ontology-based system using an AI Assistant. The aim is to simplify the structuring of historical information through a process of instance generation, allowing the transformation of archival content into formalized and semantically enriched entities —such as machines, production spaces, historical events, and actors—based on a specific information ontology for industrial heritage. This study addresses a critical gap by introducing AI-driven methodologies for semi-automation in heritage practices offering new opportunities for industrial heritage documentation and interpretation.*

1. INTRODUCTION

In the field of industrial heritage, digitalization is no longer merely the conversion of media; it has become a knowledge-oriented strategy that makes the technical knowledge embedded in archives and collections traceable, verifiable, and reusable. Foundational charters, from the Nizhny Tagil Charter to the Joint International Council on Monuments and Sites, The International Committee for the Conservation of the Industrial Heritage (ICOMOS–TICCIH) Principles for the Conservation of Industrial Heritage ("Dublin Principles"), call for comprehensive, interoperable documentation capable of linking artefacts, processes, actors, and sources over time[1-2]. On this basis, the CIDOC Conceptual Reference Model (CIDOC CRM), recognized as International Organization for Standardization (ISO) standard 21127:2023, provides the shared conceptual language for integrating

[1] TICCIH (2003). *The Nizhny Tagil Charter for the Industrial Heritage*. TICCIH.
[2] ICOMOS & TICCIH (2011). *Joint ICOMOS–TICCIH Principles for the Conservation of Industrial Heritage Sites, Structures, Areas and Landscapes* ("Dublin Principles").

heterogeneous data within a standardized semantic framework[3-4]. Nevertheless, the integration of unstructured historical texts (registers, technical correspondence, factory manuals) remains a bottleneck: advances in Handwritten Text Recognition (HTR) and Optical Character Recognition (OCR) have made transcription scalable, but the subsequent transformation of such texts into structured, ontology-aligned knowledge is still time-consuming and difficult to reproduce[5-6]. This contribution critically addresses this research gap by presenting an artificial intelligence (AI)–based workflow that combines Retrieval-Augmented Generation (RAG) to anchor outputs to their sources[7-8], schema-constrained extraction via function calling/structured outputs to generate instances consistent with an ontological[9] profile, and validation and provenance tracking, with the results published as Resource Description Framework (RDF) and queryable using the SPARQL Protocol and RDF Query Language (SPARQL). The use of linguistic models and semantic workflows allows for the rapid structuring of heterogeneous sources; however, in the field of industrial heritage, automation interacts with a complex interpretative horizon (historical terminology, multiple actors, documentary gaps). For this reason, the results are considered as assisted reading proposals, not as documentary truths. Indeed, the primary objective of this paper is to explore the potential application of AI-driven semi-automatic interpretation in instance generation, to assess the effectiveness of the results, and provide an initial evaluation of the technology's reusability in this specific context. Secondly, the aim is to highlight the present risks in terms of epistemological issues, typical source errors, and methodological gaps that need to be addressed in the near future for the practical application of the technology.

2. BACKGROUND

In the cultural heritage domain, the CIDOC Conceptual Reference Model (CIDOC CRM) and its corresponding update as International Organization for Standardization (ISO) 21127:2023 form the conceptual foundation for semantic interoperability[10-11]; building on that foundation are domain extensions such as the CIDOC CRM extension for archaeological excavations (CRMarchaeo)—focused on excavation processes, evidence, and stratigraphy—and the CIDOC CRM extension for the documentation of standing buildings (CRMba)—aimed at building archaeology—which refine the representation of processes, evidence, and spatiotemporal relationships typical of sites and historic industrial complexes[12-13]. To describe the organization of historic buildings in a minimal and

[3] ISO (2023). ISO 21127:2023 – Information and documentation – A reference ontology for the interchange of cultural heritage information.

[4] CIDOC Conceptual Reference Model (CRM). CRMbase standard provides the basic classes and relations devised for the cultural heritage world. This base ontology is complemented by a series of modular extensions to the basic model.

[5] G. MÜHLBERGER., L. SEAWARD, M. TERRAS, ET ALII, *Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study*. Journal of Documentation, 75(5), 954–976., 2019

[6] J. NOCKELS, P. GOODING, S. AMES, M. TERRAS, *Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research*. Archival Science, 22(3), 367–392. https://doi.org/10.1007/s10502-022-09397-0, 2022

[7] P. LEWIS, E. PEREZ, A. PIKTUS, ET ALII, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. In: NeurIPS 2020 (proceedings), 2020.

[8] Y. GAO, Y. XIONG, X. GAO, ET ALII, *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv:2312.10997 (v5: 27 Mar 2024), 2023.

[9] OpenAI (2024–2025). Structured model outputs – Function calling / JSON Schema con strict: true.

[10] ISO (2023). ISO 21127:2023 – Information and documentation – A reference ontology for the interchange of cultural heritage information.

[11] CIDOC Conceptual Reference Model (CRM).

[12] CIDOC CRMarchaeo — An extension of CIDOC CRM to support archaeological excavation.

composable way, the Building Topology Ontology (BOT) provides a reusable core vocabulary that is well established in the literature[14]. On the access and reusability front for digital sources, the International Image Interoperability Framework (IIIF)—in particular the Presentation Application Programming Interface (API) 3.0—and the Europeana Data Model (EDM) have consolidated practices and specifications for manifests and metadata, enabling pipelines for contextualization and content reuse[15-16]. Standards-based platforms such as Arches demonstrate the maturity of open-source solutions for the inventory and management of resources, events, actors, and sources, while ArCo—the Italian Cultural Heritage Knowledge Graph (ArCo)—exemplifies the construction of a national knowledge graph grounded in modular ontologies and conversions to Resource Description Framework (RDF) that are queryable via the SPARQL Protocol and RDF Query Language (SPARQL)[17-18-19-20-21]. In parallel, generative artificial intelligence (AI) enables linking between texts and ontologies: Retrieval-Augmented Generation (RAG) improves adherence to and verifiability against the sources[22-23]; the use of structured outputs guides extraction toward the expected classes and properties; the Shapes Constraint Language (SHACL) and the PROV Ontology (PROV-O) could ensure robustness and validation of the graphs, closing the loop in a pipeline from text to ontological instances to RDF graphs that is ready for use in the documentation and valorization of industrial heritage[24-25].

Based on previous research work in ontology engineering and Large Language Models (LLMs), the following section illustrates the proposed layered architecture approach that guides text analysis towards validating RDF.

3. AI WORKFLOW FOR INDUSTRIAL HERITAGE: METHODOLOGICAL FRAMEWORK AND CONSIDERATIONS

The use of AI, specifically OpenAI models, is considered an operational tool that speeds up access and structuring in a much larger, documental, and iterative process, where historical knowledge emerges between the model, the documents, and the researcher's expertise.

---

[13] P. RONZINO, F. NICCOLUCCI, A. FELICETTI, M. DOERR, *CRMba: a CRM extension for the documentation of standing buildings*, International Journal on Digital Libraries, 17(1), pp. 71–78, 2016. https://doi.org/10.1007/s00799-015-0160-4.

[14] M. H. RASMUSSEN, M. LEFRANÇOIS, G.F. SCHNEIDER, P. PAUWELS, *BOT: The Building Topology Ontology of the W3C Linked Building Data Group. Semantic Web Journal,* 2021. https://doi.org/10.3233/SW-200385.

[15] IIIF Consortium (2020–). IIIF Presentation API 3.0.

[16] Europeana (2016–2023). Europeana Data Model (EDM) — Documentation (Primer, Definition, Mapping Guidelines). https://pro.europeana.eu/page/edm-documentation?utm_source=chatgpt.com.

[17] D. MYERS, A. DALGITY, I. AVRAMIDES, *The Arches heritage inventory and management system: a platform for the heritage field. Journal of Cultural Heritage Management and Sustainable Development*, 6(2), 2016, pp. 213–224. (PDF Getty).

[18] Getty Conservation Institute (s.d.). Arches — Project overview.
https://www.getty.edu/projects/arches/?utm_source=chatgpt.com

[19] 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: ESWC 2019 Satellite Events (LNCS). https://doi.org/10.1007/978-3-030-30796-7_3

[20] W3C (2014). RDF 1.1 Concepts and Abstract Syntax

[21] W3C (2013). SPARQL 1.1 Query Language

[22] P. LEWIS, E. PEREZ, A. PIKTUS, et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. In: NeurIPS 2020 (proceedings).

[23] Y. GAO, Y. XIONG, X. GAO, et al, *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv:2312.10997, 2023 (v5: 27 Mar 2024).

[24] W3C (2017). Shapes Constraint Language (SHACL).

[25] W3C (2013). PROV-O: The PROV Ontology.

The proposed system architecture is organized in four layers:

- Knowledge & data sources: unstructured archival documents (reports, registers, drawings).
- Ontology layer (IndArch): the domain schema, covering machines, spaces, processes, events, and their relations, acts as a contract for extraction and typing.
- AI Assistant layer: a backend service ingests files, creates embeddings, and indexes them in a vector store; a retriever (RAG) selects relevant passages, which are interpreted by the LLM under ontology-guided constraints. Outputs are enforced through a function-calling schema aligned with IndArch (allowed classes, object/data properties).
- Knowledge formalization layer: the Assistant returns JSON instances compliant with the schema; a Google Colab script in Python produces RDF/Turtle and publishes to a triplestore/Protégé. This layered design separates storage, reasoning, and validation, ensuring that only ontology-conformant instances are added to the knowledge base.

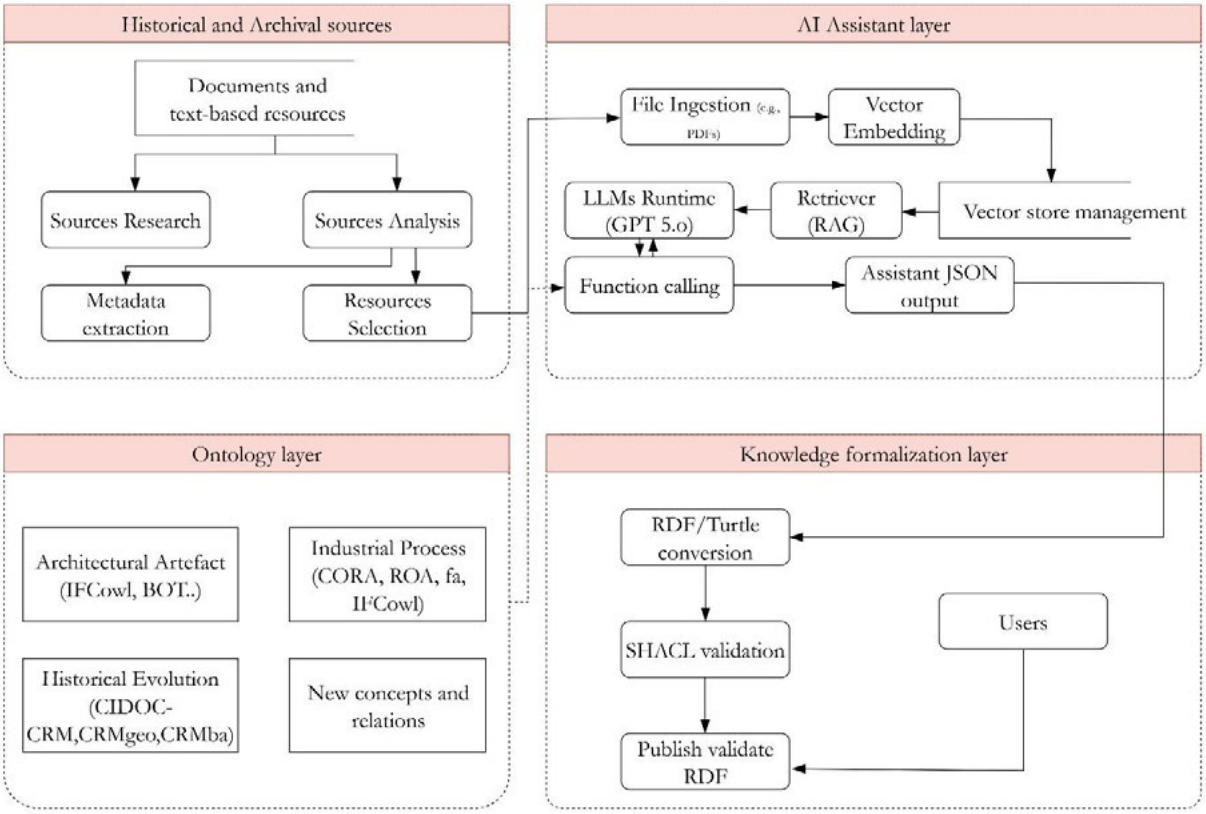

Fig. 1. System architecture of the AI workflow for industrial heritage.

The application workflow (Fig. 2) for this paper targets a subset of classes, object properties, and data properties of the Indarch Ontology in order to validate the defined process. A function calling schema was defined that aligns with the ontology structure. The written document's information is extracted from the text based on the function, and the output is in JSON format. At this stage, the post-processing of the JSON file into RDF format was performed on the Google Colab platform, which runs on a cloud server. The output of this post-processing is in RDF format, aligned with the ontological structure defined in the function calling. The resulting RDF can be opened in any RDF tool or triplestore (e.g., Protégé) for querying and further analysis. The final result can be compared

to the one that was performed manually in previous research, to start tracing initial comments and discussion on the behavior of the model.

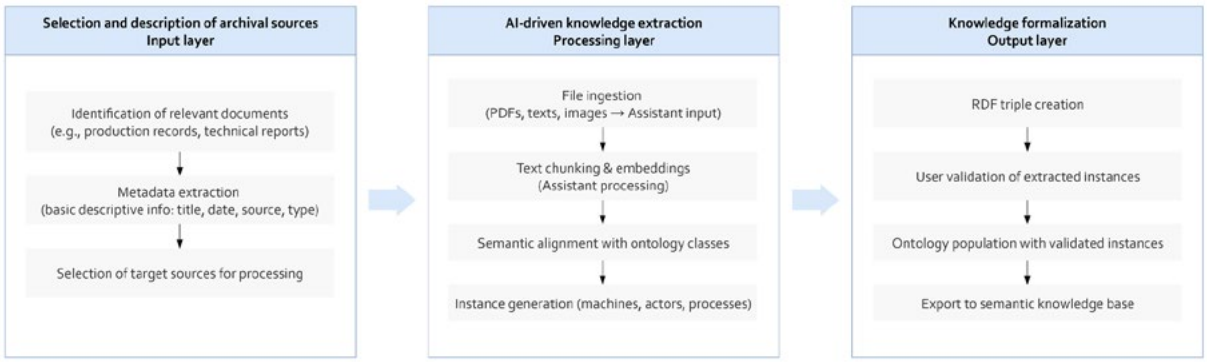

Fig. 2. Application workflow.

4. EXPERIMENTAL APPLICATION AND DISCUSSION

The experimental application regards the instantiation of the historical paper production process phases of the papermills located in the Tivoli area. In this first stage the authors considered three written sources:

- "The modern Paper-Making Machine". Author: W.H.Orr.
- "Storia dei materiali scrittori e delle forme del libro: un'introduzione 1 . La carta e la tapa". Author: Carlo Pastena
- "Carta e stracci. Protoindustria e mercati nello Stato pontificio tra Sette e Ottocento". Author: Augusto Ciuffetti.

The small subset of classes, object properties and data properties considered is defined as follows:
Classes

- *fa:ManufacturingTask*
- *indarch:HumanLabour*
- *indarch:MachineLabour*
- *fa:CapabilityDescr*
- *ifc:IfcSpace*
- *indarch:Machine*

Object properties (domain → range)

- *fa:hasComponentReq*: *fa:ManufacturingTask* → *cidoc:E18_PhysicalThing* (materials/parts required)
- *fa:hasResourcesReq*: *fa:ManufacturingTask* → *fa:CapabilityDescr* (skills/capabilities required)
- *fa:hasOutput*: *fa:ManufacturingTask* → *cidoc:E18_PhysicalThing*
- *fa:Satisfies*: (*indarch:HumanLabour* ∪ *indarch:MachineLabour*) → *fa:CapabilityDescr*
- *indarch:isPerformedIn*: *fa:ManufacturingTask* → *ifc:IfcSpace (room/area)*

Data properties for indarch:Machine

- *indarch:Manufacturer* (string)
- *indarch:Model* (string)
- *indarch:YearOfManufacture* (string or xsd:gYear)
- *indarch:TechnicalSpecification* (string)

The prompt sent to the AI Assistant in the OpenAI API playground is:

From all files, extract a small, clean graph of papermaking tasks, their inputs/outputs, and any named machines. Keep labels concise; ensure relation targets align with existing labels. Top-level provenance: "Multiple sources"; add per-instance sources (filenames). Flag uncertainties with requiresReview: true + note. Use the function only.
Based on the defined function in JSON, the following is an extraction of the resulting output from this run of the AI Assistant.

```
{
  "@class": "fa:ManufacturingTask",
  "label": "Collation and Pressing",
  "relations": [
    {"p": "fa:hasComponentReq", "o": "Glue Mixture"},
    {"p": "indarch:isPerformedIn", "o": "Press Section"}
  ],
  "sources": ["Pastena, 2018, Storia dei materiali scrittori.pdf"],
  "requiresReview": True,
  "note": "Exact specifications of inputs/outputs are unclear from the text."
}
```

In this part of the extraction, for instance, the specifications of the input and output were unclear to the Assistant; therefore, it specifically requested a review from the expert to check the text directly.
The expert can compare it with the cited passage and correct the label; the correction is directly saved back in the RDF, in this case, in the ontology editor Protégé.
In the same run, other examples included the pulp beating task and the forming of the paper sheet. After that, the JSON was pasted into the Colab notebook, which first adds a minimal TBox to display classes and object properties, then serializes it to RDF/Turtle triples per instance (fig. 3).

These initial results demonstrate how the combination of a tight function schema and concise labels can produce instances without overgeneralization by attaching filenames in the sources and explicitly requesting reviews when necessary. The interoperable output can be used as a basis for further analysis and reasoning and adopted in other information systems.
Typical issues observed in these runs of the AI Assistant are the granularity mismatch; in fact, the model sometimes proposes broad tasks (e.g., papermaking). Historical terminology variance is another issue that requires glossary alignment with specific vocabulary. Furthermore, entity span noise causes the model to extract partial or overlapping items unless guided by examples.
In a broader context, the implications of the experimental results show that the output is standard RDF/Turtle with stable identifiers, allowing it to be opened directly in Protégé or any triplestore and queried alongside other datasets. This interoperability enables graphs from different archives to be merged and compared, making recurring patterns (tasks, inputs/outputs, spaces, machines) visible across case studies. Moreover, the same data can be reused in other systems (e.g., registries, HBIM/GIS links) with minimal additional effort.

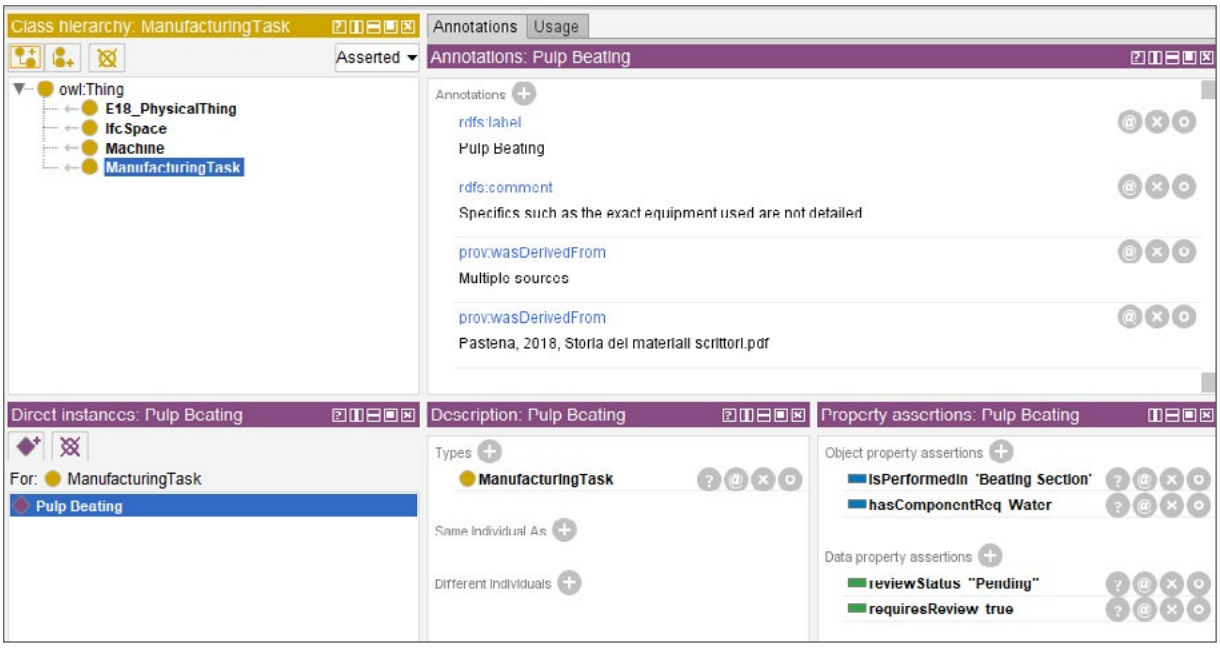

Fig. 3. Example of extracted instances visualized inside the Protégé software, aligned with the ontology schema.

5. CONCLUSIONS

This research proposes an ontology-guided AI workflow that converts unstructured text into semantic knowledge for industrial heritage. The framework utilizes Retrieval Augmented Generation (RAG), schema-constrained extraction (utilizing function calls with a restricted ontology scope), and RDF/Turtle output, thereby shifting the effort towards expert validation and interpretation.

In future works, scaled across many documents, the approach could enable transversal reading of historical production processes: instead of dealing with each site as an isolated case, recurring patterns, tasks, inputs/outputs, spaces, and machines emerge, supporting a broader, comparative view of industrial heritage.

Furthermore, future experiments and implementations will focus on providing the model with more contextual knowledge, including few-shot learning, domain exemplars, terminology guides (such as mini-glossaries), reranking, and filters to mitigate noise and hallucinations, while maintaining the tagging of instances that require expert review.

The integration of AI workflows in research and documentation activities offers novel knowledge management pathways for industrial heritage, leading to more informed reuse and valorization processes.