Towards a Knowledge-Graph-Driven Retrieval-Augmented Generation for Exploring and Curating Active Archives

Danae Pla Karidi[1], Christos Chrysanthopoulos[2] and Ioannis Triantafyllou[2]

[1] *Archimedes, Athena Research Center, Athens, Greece*
[2] *University of West Attica, Athens, Greece*

Abstract

Active archives are dynamic collections that continuously grow and evolve through continuous ingestion and metadata enrichment. This vision paper outlines a modular architecture that fuses semantic metadata management with knowledge-graph-driven (KG-driven) exploration and retrieval-augmented generation (RAG) data curation. In our design, an exploration pipeline leverages domain ontologies and knowledge graphs to help users refine their queries and discover relevant information. Complementing this, a RAG-enabled curation pipeline combines retrieved archival content with generative AI, specifically large language models (LLMs), to synthesize and summarize findings into coherent narratives. We pose three research questions on retrieval quality, annotation accuracy, and usability for future evaluation. The framework is domain-agnostic and can be applied to any digital archive or library collection.

Keywords

Retrieval-Augmented Generation, Knowledge Graphs, Active Archives, Data Exploration, Automated Curation

1. Introduction

Generative AI systems, powered by large language models (LLMs), are transforming how we interact with complex information. Digital libraries and large digital archives now act as active archives. Archives hold unique records with provenance, libraries published works, but both keep growing and need ongoing curation. Yet in digital archives, non-experts still struggle to locate context, while archivists labour to maintain consistent semantic metadata, forming a discovery-curation challenge as collections grow. Traditional keyword interfaces reveal only fragments of data, limiting broader context. Research on generous interfaces [1] and faceted browsing [2] underscores the need for knowledge-driven exploration tools. We define an *active archive* as a born-digital or digitized collection that ingests new content continuously, requiring ongoing curation. Even hybrid archives that combine physical and digital records can exhibit this active, evolving behavior. Therefore, maintaining large active archives requires continuous curation, enriching documents with metadata and tags to keep the knowledge well-structured. In this vision paper, we propose a modular knowledge-based architecture with two connected components: (a) a user-facing knowledge-guided exploration system, and (b) an expert-facing retrieval-augmented generation (RAG) curation system. A key strength of this architecture is the dynamic feedback loop between exploration and curation. When a particular topic or query is searched frequently by users, the system flags that interest for archivists, who can then add or refine metadata. Updates appear in user-facing exploration in real-time, ensuring the archive evolves alongside engagement. Our method fuses discovery and curation in real time, creating a living archive. To our knowledge, this is the first work to couple KG-guided exploration with RAG-based curation in a single continuous loop. Both modules share a unified knowledge infrastructure (e.g., DBpedia [3] and internal archives) for a consistent factual base. We formulate three research questions: (1) RQ1 (Exploration): Does KG-guided keyword expansion improve top-k precision for non-expert queries?, (2) RQ2 (Curation): How accurate are RAG-generated metadata suggestions after curator review?, (3) RQ3

SEMANTiCS 2025: 21st International Conference on Semantic Systems, September 03 - 05, 2025, Vienna, AT

*Corresponding author.
danae@athenarc.gr (D. Pla Karidi)
0000-0002-3154-6212 (D. Pla Karidi)

(Usability): What usability barriers emerge in a semantic-metadata interface for active archives?

2. Related Work

In this section, we briefly summarize relevant prior research on active data, knowledge-based interfaces, and retrieval-augmented generation in digital archives and libraries. **Active Data** denotes datasets still being generated, processed, or iteratively refined. "Active curation" locates curatorial work at the very start of the data lifecycle, well before analysis or publication. Early technical, organisational, and human choices determine usability and long-term value, yet coherent, integrated guidance is scarce. Collaboration between researchers and data professionals is therefore essential at the point of creation [4]. In Computational Archival Science (CAS), the "Vanishing Box" [5] signals the collapse of orderly structure in digital records: messy, fragmented objects require generative AI, knowledge graphs, and other novel methods to restore meaning. Active curation is both technical and interpretive, ensuring today's complex records remain accessible, comprehensible, and usable [5]. Lifecycle models cast curation as a chain of stages, from planning and collection to preservation, reuse, and eventual disposal, aimed at maximising value over time [6]. Empirical work reinforces this view [7]. Active curation is equally central to open science: a pragmatic, incremental stance maintains that partial reproducibility and accessibility beat inaction, urging curators to engage as data are produced, even with limited expertise, to enhance later discoverability and reuse [8]. Consequently, digital library collections function as *active archives*, growing continuously and demanding constant metadata maintenance. **Knowledge-Based Interfaces for Non-Expert Users in Digital Archives and Libraries.** Traditional keyword search often frustrates non-experts, the public or humanities scholars unfamiliar with archival structure, because a lone search box presumes knowledge of both collection and terminology, an "ungenerous" design that withholds context [1]. To lower this barrier, digital-heritage research promotes exploratory, knowledge-based interfaces. Faceted search, long established, exposes structured metadata (dates, places, topics) so users filter results incrementally, cutting zero-hit queries and teaching archive vocabulary [2]. Rich-prospect or "generous" views go further: thumbnail grids, concept maps, and other overviews foster browsing and serendipity without assuming domain expertise, instead turning metadata and contextual links into navigation aids [1]. Increasingly, such interfaces draw on knowledge graphs [9]. For example, [10] visualises entities (people, places, works) and their relationships, letting users traverse semantic links rather than isolated records. Knowledge-based UIs thus allow non-experts to assemble broader narratives, showing how ontologies, thesauri, and graphs underpin generous, concept-centric exploration. **Retrieval-Augmented Generation for Metadata Curation in Digital Archives.** Rich metadata powers sophisticated archive interfaces but is expensive to produce [11]. Recent work exploits large language models (LLMs) to draft or complete records with minimal supervision. Song et al.[12] cast description assignment as zero-shot classification, while Huang et al.[13] prompt GPT-4 to generate titles and abstracts for archived pages. ChatGPT cataloguing pilots also report time savings, though curator review remains essential. Retrieval-Augmented Generation (RAG) limits hallucinations by retrieving evidence before conditioning the LLM output [14]. Nguyen et al. [15] show that a hybrid RAG pipeline boosts answer relevance and surfaces citations in an archival search prototype. Overall, coupling semantic retrieval with LLMs enhances user access and back-office metadata quality. On another front, general-purpose library platforms, like Europeana [16] and the Digital Public Library of America (DPLA) [17], enrich metadata offline and rely on manual provider updates; no mechanism links live user queries to immediate, AI-assisted curation, highlighting the gap of automated curation solutions that enable continuous metadata injection.

3. System Overview

We propose a modular architecture in which knowledge-based exploration and RAG-based curation share a common knowledge layer: external graphs, an internal KB, and archive metadata, ensuring factuality, adaptability, and consistency. Because both pipelines draw on the same grounding facts, each

LLM-powered module can evolve independently while their outputs reinforce one another: exploration reveals gaps for curators, and curator-approved updates enrich exploration in real-time. This closed loop minimizes hallucinations, builds trust, and keeps the archive current (Fig. 1).

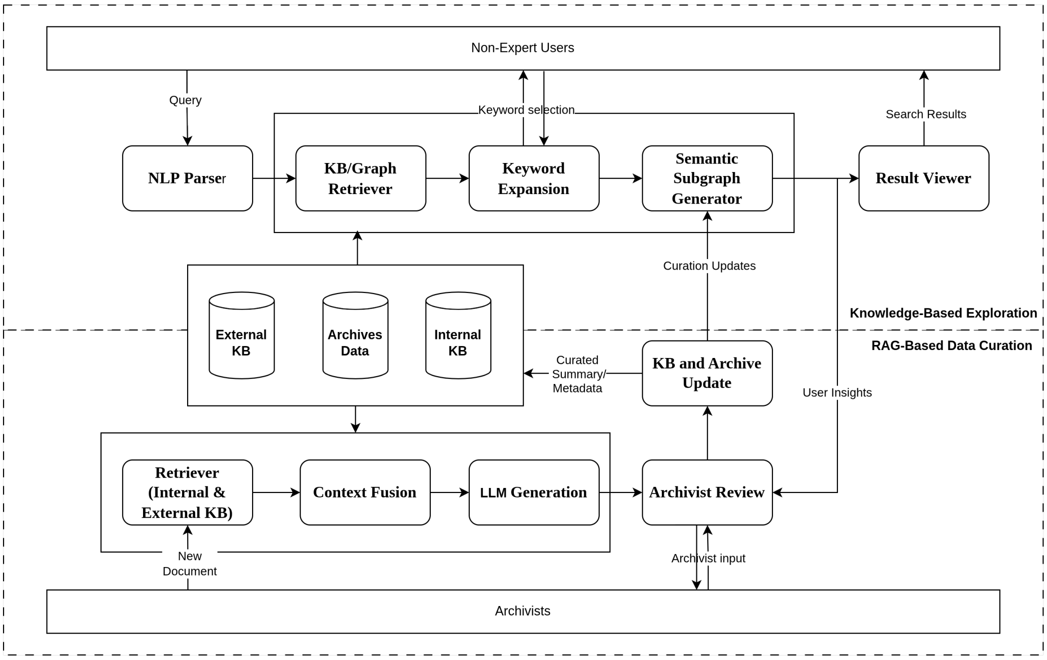

Figure 1: The unified exploration and curation framework architecture.

3.1. Knowledge-Based Exploration for Non-Expert Users

For exploration, the system enables users to navigate the archive in natural language, augmented by intelligent suggestions. Instead of formulating precise queries or browsing static categories, a non-expert user can ask a question or type a broad topic, and the system will guide them by leveraging background knowledge. **User Input:** The user provides a natural-language query (or a short list of keywords), using their informal vocabulary. **NLP Parser:** detects key entities and broader topics in the user query, then assigns each to a node in the knowledge graph (e.g. DBpedia) with a confidence score. **KB / Graph Retriever:** receives the parser frame and retrieves relevant nodes and relations from three sources: (i) external public KGs, (ii) the internal domain KB, and (iii) the archive's own metadata and full-text index. Then merges the results into a focused heterogeneous subgraph centered on the seed entities. This subgraph bridges the gap between the non-expert user's natural-language query and specialized domain concepts. **Keyword Expansion:** recommends additional keywords or phrases to narrow/widen or align the search scope to the existing internal data. To do this, it mines the generated subgraph using graph mining techniques (co-occurrence, embedding similarity, PageRank centrality, etc.) to select up to k top-relevant alternative nodes to recommend to the user. The resulting keyword graph highlights central nodes and presents them so users can adopt more precise terminology or explore related topics. **Semantic Subgraph Generator:** after the user selects some of the recommended keywords, this module renders an interactive subgraph that shows the user's query alongside related concepts, entities, and linked archive items. New metadata produced by the RAG pipeline, new tags, is incorporated instantly, while heavily travelled paths or recurring searches are flagged for curators to expose gaps or emerging themes. This closed-loop mechanism keeps the subgraph and the archive evolving in sync with user interests. **Result Viewer:** renders (i) a ranked list of documents with snippet highlights, (ii) an interactive graph pane showing the semantic subgraph, and (iii) dynamically generated facets (entities, time, location, topic). User interactions: node click, facet toggle, feedback thumbs emit events logged for both analytics and online learning. Users see not just a list of items but also how each relates to broader concepts. All relevance scores are exposed to support explainability. By grounding recommendations in a knowledge graph, the system keeps them understandable and context-aware, lowering barriers for exploring complex archives. It extends the notion of faceted search by dynamically generating AI-driven facets for each query instead of relying on static categories alone. This context-rich interface provides multiple pathways for discovery, revealing how information is interlinked across the archive. The continuous feedback loop, which synchronizes user exploration with metadata updates, ensures the archive remains current, comprehensive, and responsive to emerging interests.

3.2. RAG-Based Approach for Active Data Curation

The RAG pipeline curates new documents under human oversight. As new documents arrive, the pipeline generates metadata, summaries, and knowledge base entries under human supervision. The objective is to enrich and maintain the archive's structured knowledge using AI efficiently, while a human expert ensures quality control. **New Document Input:** accepts single files or batches and the curator initiates the pipeline rather than annotating it manually. The system ingests the document (or batch) as input, ready for automated processing. **Retriever (Internal & External):** for each incoming document, queries two retrieval tiers: (1) semantic retrieval over the existing archive to find similar contents, and (2) KG-aware retrievers on external sources (e.g. DBPedia) to fetch definitions and reference facts for every mentioned entity. **Context Fusion:** builds a composite prompt comprising the new document's content with supporting facts. Relevant excerpts are concatenated and organized (e.g., by topic), ensuring the LLM has grounded information, for instance, pairing an external knowledge-base definition of each entity with the sentence where it is mentioned in the new document. **LLM Generation:** processes the fused context and the document's content to generate the desired outputs. This can include a draft summary of the document, a set of suggested metadata tags, identified entities (with potential links to the knowledge graph), and even relational triples (for updating the graph). Crucially, because the LLM has been fed with relevant context, these outputs are grounded, allowing the model to cite or incorporate actual facts from the retrieved documents instead of hallucinating. **Archivist Review:** inspects the LLM's outputs for accuracy, relevance, and proper sourcing, correcting or confirming as needed. Because suggestions are backed by references, each piece of metadata is traceable to its source. After approval, new or revised metadata updates the knowledge graph and indexes. This human-in-the-loop step ensures curation quality. **KB & Archive Update:** integrates approved metadata and new knowledge into the archive's internal knowledge base, updating the index, tags, and relationships. A scheduled offline job learns from archivist edits, refining prompts and retrieval to improve accuracy. Knowledge-based exploration and Active Curation can operate separately or integrate to share insights in real-time. User exploration logs (such as frequent unanswered queries or clicked concepts) inform the curators about emerging interests or gaps in the archive, prompting targeted curation. while newly curated metadata instantly appears in the exploration interface. This continuous loop ensures an evolving, living archive: documents are constantly updated and made discoverable through both AI automation and expert oversight.

4. Conclusion and Future Work

Answering the posed research questions will guide our future evaluation of the system using both expert judgment and quantitative metrics. We plan to evaluate whether KG-guided keyword expansion improves top-*k* precision (RQ1) by comparing search results with and without the expansion on a set of test queries (e.g. using precision@k and recall metrics). To assess RAG-generated metadata accuracy (RQ2), we will conduct a study where archivists review AI-suggested tags/ summaries. We will measure the acceptance rate of suggestions and the frequency of corrections needed, thereby evaluating accuracy after human review. We will also compare the factual consistency against the source (e.g. using metrics like FactScore[18], hallucination rate = incorrect factual statements / total statements) against a zero-shot variant without retrieval. Usability (RQ3) will be evaluated through user studies:

we will gather feedback from non-experts and archivists using the system. Metrics like task success rate, user satisfaction, and qualitative feedback on the interface will help identify any usability barriers. **Declaration on Generative AI:** The authors have not employed any Generative AI tools.

Acknowledgments

The publication of this article was funded by the University of West Attica.

References

[1] M. Whitelaw, Generous interfaces for digital cultural collections, Digital humanities quarterly 9 (2015) 1–16.

[2] K.-P. Yee, K. Swearingen, K. Li, M. Hearst, Faceted metadata for image search and browsing, in: Proceedings of the SIGCHI conference on Human factors in computing systems, 2003, pp. 401–408.

[3] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al., Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia, Semantic web 6 (2015) 167–195.

[4] I. Kouper, K. L. Tucker, K. Tharp, M. E. van Booven, A. Clark, Active curation of large longitudinal surveys: A case study, Journal of EScience Librarianship 10 (2021).

[5] J. Proctor, R. Marciano, A computational review of the literature of computational archival science (cas): Advancing archival theory in the age of the digital tsunami and the vanishing box problem, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 2514–2523.

[6] G. Oliver, R. Harvey, Digital curation, American Library Association, 2016.

[7] H. L. Rhee, A new lifecycle model enabling optimal digital curation, Journal of librarianship and information science 56 (2024) 241–266.

[8] S. L. Sawchuk, S. Khair, Computational reproducibility: A practical framework for data curators, Journal of eScience Librarianship 10 (2021).

[9] D. Pla Karidi, Y. Stavrakas, Y. Vassiliou, Tweet and followee personalized recommendations based on knowledge graphs, Journal of Ambient Intelligence and Humanized Computing 9 (2018) 2035–2049.

[10] C. S. Khoo, E. A. Tan, S.-G. Ng, C.-F. Chan, M. Stanley-Baker, W.-N. Cheng, Knowledge graph visualization interface for digital heritage collections, Information Technology and Libraries (2024).

[11] I. Triantafyllou, Thematic categorization on university records, in: 2023 IEEE 11th International Conference on Systems and Control (ICSC), IEEE, 2023, pp. 384–389.

[12] H. Song, S. Bethard, A. Thomer, Metadata enhancement using large language models, in: Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), 2024, pp. 145–154.

[13] A. Y. Huang, A. Nair, Z. R. Goh, T. Liu, Web archives metadata generation with gpt-4o: Challenges and insights, arXiv preprint arXiv:2411.05409 (2024).

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[15] H. D. Nguyen, T.-H. A. Nguyen, T. B. Nguyen, A proposed large language model-based smart search for archive system, in: International Symposium on Information and Communication Technology, Springer, 2024, pp. 210–223.

[16] B. Haslhofer, A. Isaac, data.europeana.eu – the europeana linked open data pilot, in: Proc. DC-2011, 2011.

[17] S. Bragg, S. Tumlin, Metadata aggregation at the digital public library of america, in: Proc. JCDL 2015, 2015.

[18] S. Min, P. Lewis, W.-t. Yih, Factscore: Evaluating factual consistency in retrieval-augmented generation, in: Proc. ACL 2023, 2023.