DCPAPERS

DCMI 2026

Proceedings of the International Conference on Dublin Core and Metadata Applications 2026

ISSN: 1939-1366

AI-Guided Metadata Construction for Meaning-Driven Digital Knowledge Systems: A Framework for Automated Metadata Generation and Semantic Discovery

Wirapong Chansanam[1], **Umawadee Detthamrong**[1], **Chunqiu Li**[2], **Abdul Rahman Ahmad**[3], and **Avshalom Elmalech**[4]

[1]Khon Kaen University, [2]Beijing Normal University, [3]Universiti Teknologi Mara, [4]Bar-Ilan University

Abstract

The rapid expansion of digital repositories and scholarly resources has increased the demand for scalable and intelligent metadata management systems. Traditional metadata creation methods, which rely on manual cataloguing by information professionals, struggle to keep pace with the growing volume and heterogeneity of digital content. This study develops and evaluates an AI-guided framework, Metadata-Building-AI-Guidance V.1.0 that combines document ingestion, chunking, OpenAI text-embedding-3-small embeddings, and the gpt-4o-mini large language model into a single Streamlit application supporting both automated Dublin Core-aligned metadata extraction and embedding-based semantic retrieval. We position the system as a design-science artefact in which retrieval is not a side feature but a feedback loop: the same vector index that powers similarity search is also used to surface the contextual chunks from which structured metadata are extracted and to support retrieval-augmented question answering over uploaded collections. The system was evaluated on a purposively sampled corpus of 30 open-access academic documents and 20 expert queries, using (i) field-level F1 against librarian-curated ground truth with Cohen's κ for inter-annotator agreement and (ii) Precision@5 and Mean Reciprocal Rank with binary relevance judgements from two librarians. Field-level F1 ranged from 0.67 to 0.73 for dc:title, dc:creator, dc:subject, and dc:description, with overall κ = 0.83 (almost perfect agreement); the lowest F1 was 0.57 for dc:type, traced to a fixed generic prompt output rather than to a model-capability limitation. Semantic retrieval reached Precision@5 = 0.61 and MRR = 0.69 with κ = 0.66 (substantial agreement) across the 20 queries. We discuss the limits of LLM-only evaluation—including the absence of a head-to-head comparison with established non-LLM extractors such as GROBID—and identify controlled baseline comparison together with a refined dc:type prompt as the immediate next steps. Prompts, JSON schema, library versions, sampling log, and evaluation queries are released to support replication. The contribution is a reproducible reference implementation that aligns AI-assisted metadata extraction with Dublin Core Terms and the FAIR principles for digital libraries, archives, and cultural-heritage repositories.

Full Paper

Available Aug 01, 2026

Correspondence to
Wirapong Chansanam
wirach@kku.ac.th

1. Introduction

In the era of digital scholarship, the volume of digital information is expanding at an unprecedented rate. Academic repositories, digital libraries, and archival systems collectively manage millions of documents, datasets, and multimedia objects, creating significant challenges for organizing, describing, and retrieving knowledge effectively. Despite advances in digital infrastructure, metadata creation remains largely manual and labor-intensive. This situation raises an important question: How can libraries and information systems scale metadata creation to keep pace with the exponential growth of digital resources?

Metadata plays a critical role in enabling resource discovery, interoperability, and long-term preservation within knowledge organization systems. Traditionally, meta-

data creation has relied on manual cataloging performed by librarians and domain experts. While this process ensures high-quality descriptive information, it is time-consuming and difficult to scale in large digital collections. Consequently, scholars have increasingly explored the integration of Artificial Intelligence (AI), Natural Language Processing (NLP), and machine learning techniques to automate metadata generation and enrichment processes. Recent systematic reviews demonstrate that AI technologies can significantly enhance metadata workflows by automating classification, entity extraction, and semantic enrichment [1], [2]. Similarly, studies on AI-enabled cataloging indicate that machine learning methods can improve the efficiency and consistency of metadata creation while reducing the workload of library professionals [3], [4].

Furthermore, research across library science and digital repository management highlights the growing adoption of AI-driven metadata extraction techniques. NLP-based approaches have been applied to automate indexing, classification, and document annotation, enabling more efficient management of large volumes of scholarly content [5], [6]. In digital preservation and institutional repositories, AI-assisted semantic enrichment has been shown to improve information retrieval by linking concepts, entities, and contextual knowledge within metadata records [7]. Similarly, deep learning approaches have demonstrated promising capabilities in organizing multimodal archival resources, including image categorization and content-based retrieval [8]. These developments suggest that AI technologies have the potential to transform traditional metadata workflows into intelligent, scalable knowledge organization systems.

However, despite these promising advances, several challenges remain. Systematic reviews consistently report issues related to data quality, ethical considerations, and technical limitations in AI-based metadata systems [9], [10]. In addition, researchers emphasize the need for standardized frameworks and robust metadata structures to support interoperability across heterogeneous digital environments [11]. Without clear metadata standards and governance mechanisms, automated systems may produce inconsistent or unreliable descriptions, undermining the reliability of digital collections.

Another important limitation concerns the practical implementation of AI-assisted metadata systems. While many studies highlight the theoretical potential of AI technologies, fewer works provide concrete system architectures or operational frameworks for integrating AI guidance into real-world metadata workflows [12], [13].

Additionally, emerging research on data catalogs and metadata integration demonstrates the growing importance of AI-driven metadata ecosystems. AI-based approaches are increasingly used to support ontology mapping, semantic alignment, and automated concept identification across heterogeneous datasets [14], [15]. Similar trends are observed in archival science and records management, where AI techniques have been applied to automate categorization, metadata enrichment, and information extraction in digital archives [16], [17]. In library information systems, computer vision and multimodal AI technologies are also enabling new forms of metadata enrichment for digital collections, including automated annotations and descriptive labeling [18].

Although these studies collectively demonstrate the transformative potential of AI in metadata management, a critical gap remains in the development of integrated frameworks that guide the process of building metadata using AI-assisted workflows. Most existing studies focus either on conceptual discussions or on specific AI techniques, leaving limited guidance for designing practical systems that combine document processing, semantic embeddings, metadata extraction, and knowledge retrieval. Moreover, research rarely addresses how AI-generated metadata can support interactive knowledge exploration and semantic discovery, which are increasingly important in modern digital scholarship environments.

To address this gap, this study develops an AI-guided framework that treats automated metadata generation and semantic retrieval not as two parallel features but as a single integrated pipeline. In the proposed design, the same chunked, embedded representation of each uploaded document drives three coupled functions: (i) it provides the context window from which structured Dublin Core-aligned metadata are extracted by a large language model, (ii) it indexes the document for cosine-similarity retrieval at query time, and (iii) it supplies the top-k chunks used by a retrieval-augmented generation (RAG) module for question answering over the collection [19]. This shared representation is the mechanism by which retrieval directly supports metadata management: relevant passages identified through embedding search are passed back to the extraction prompt to refine or expand structured fields (for example, to recover keywords or named entities that a single-pass first-page extraction would miss), and the metadata records produced are themselves used as retrievable units. The study therefore investigates how natural language processing, semantic embeddings, and large language models can be coupled into a unified, reproducible system that produces standards-aligned metadata while remaining queryable end to end.

The significance of this research lies in its contribution to both theoretical and practical advancements in knowledge organization systems. Theoretically, the study advances understanding of how AI techniques can be integrated with semantic metadata frameworks to support scalable knowledge organization. Practically, the proposed system demonstrates how AI guidance can enhance metadata creation workflows, improve resource discoverability, and support intelligent information retrieval within digital repositories. By bridging the gap between conceptual research and system implementation, this study provides a foundation for next-generation AI-assisted metadata systems in digital libraries, archives, and data-driven research environments.

2. Methodology

2.1. Research Design

This study adopts a design science research (DSR) approach to develop and evaluate an AI-assisted metadata generation system for digital documents. Design science is appropriate because the research aims to create and evaluate a technological artifact that improves metadata production in digital collections. The artifact developed in this

study is a Metadata-Building AI Guidance platform that integrates document processing, semantic embeddings, retrieval mechanisms, and large language models (LLMs) to support automated metadata creation and semantic querying. The methodological process consists of four main stages:

1. Document ingestion and preprocessing
2. Semantic representation through embeddings
3. AI-assisted metadata extraction
4. Semantic retrieval and question answering

These stages form an integrated pipeline enabling automated metadata construction and knowledge discovery from heterogeneous digital documents.

2.2. System Architecture

The proposed system architecture consists of five main components:

1. Document ingestion layer
2. Text preprocessing and chunking module
3. Embedding and semantic indexing module
4. AI-driven metadata extraction module
5. Semantic retrieval and question-answering interface

The system was implemented using Python and Streamlit, enabling interactive document upload, metadata generation, and semantic exploration. The application processes multiple document formats including PDF, DOCX, CSV, JSON, and plain text files, which are automatically converted into machine-readable text during the ingestion stage.

The architecture enables a workflow in which uploaded documents are converted into structured textual data, segmented into manageable units such as chunks, transformed into semantic vectors, and analyzed using AI models to generate descriptive metadata.

2.3. Data Acquisition and Document Processing

The system accepts multiple document formats commonly used in digital libraries and research repositories. During ingestion, each document undergoes automated text extraction using format-specific parsing techniques.

PDF files are processed using PDF readers, while DOCX files are parsed through structured paragraph extraction. Tabular formats such as CSV and hierarchical formats such as JSON are converted into textual representations to ensure compatibility with the metadata generation pipeline.

This preprocessing stage ensures that heterogeneous document formats are normalized into a unified textual representation suitable for downstream natural language processing tasks.

2.4. Text Segmentation and Chunking

To enable efficient semantic analysis, long documents are divided into smaller segments using a token-based chunking strategy. Each document is split into segments of approximately 800 tokens with a 200-token overlap between adjacent segments. This overlapping approach preserves contextual continuity while preventing loss of semantic information at segment boundaries. Chunking serves two important purposes:

1. Improving embedding quality, since most embedding models perform better on moderately sized text segments.
2. Enabling granular retrieval, allowing the system to identify relevant sections of a document rather than processing entire documents at once.

   Each generated chunk is stored in a temporary document repository for embedding and retrieval processes.

2.5. Semantic Embedding Generation

To capture the semantic meaning of textual segments, each chunk is transformed into a dense vector representation using a neural embedding model. Specifically, the system employs the text-embedding-3-small model, which produces numerical vectors representing semantic relationships among textual units.

For each text chunk c_i, an embedding vector e_i is generated:

$$e_i = EmbeddingModel(c_i)$$

The resulting embedding vectors are stored in a vector matrix and used to construct a semantic index. This representation allows the system to measure semantic similarity between user queries and document segments using cosine similarity. Embedding vectors are stored using NumPy arrays to enable efficient vector computations and similarity calculations.

The core innovation of the system lies in its AI-guided metadata generation module, which uses the large language model gpt-4o-mini (accessed through the OpenAI Chat Completions API at temperature 0.2, top_p = 1.0, max_tokens = 1500, response_format = json_object) to analyse document content and return structured metadata. For each document, the top-k embedding-ranked chunks (k = 8) are concatenated and passed to the model together with a fixed system prompt and a JSON schema that enforces the following Dublin Core-aligned fields:

- Title
- Authors
- Keywords
- Summary
- Named entities

  The metadata extraction process is guided through structured prompting that instructs the AI model to analyze document content and return standardized

metadata fields. This approach enables the automated creation of descriptive metadata that can support digital library indexing, information retrieval, and knowledge graph construction.

2.5.1. *Prompt Template and JSON Schema*

To support reproducibility, the full extraction prompt, JSON schema, model identifier, and library versions used in this study are released alongside the manuscript at https://github.com/wirapong/Metadata-Building-AI-Guidance. The prompt instructs the model to act as a cataloguing assistant, to read only the supplied context, and to return strictly valid JSON conforming to the schema; the schema maps directly to the Dublin Core Terms elements dc:title, dc:creator, dc:subject, dc:description, and dc:type, with an auxiliary named_entities array used for downstream knowledge-graph linkage. Field values are post-validated against the schema, and malformed responses are retried up to two times before being marked as extraction failures (counted as false negatives in the evaluation below).

2.6. Semantic Retrieval Mechanism

To retrieve relevant information from indexed documents, the system implements a vector similarity search mechanism. When a user submits a query, the system generates a query embedding and compares it with the stored document embeddings using cosine similarity.

The similarity score between a query vector q and document vector d_i is computed as:

$$Similarity(q, d_i) = \frac{q \bullet d_i}{\|q\|\|d_i\|}$$

The top-(k) most similar document segments are retrieved and concatenated to form the contextual input for downstream AI processing. This retrieval strategy allows the system to focus only on the most semantically relevant document fragments, improving both efficiency and response quality.

2.7. AI-Supported Question Answering

In addition to metadata generation, the system includes a semantic question-answering module. After retrieving relevant document segments, the system sends the contextual information along with the user query to the language model, which generates a natural language response. This capability transforms the platform from a simple metadata generator into a knowledge exploration interface, allowing users to interact with document collections through conversational queries.

2.8. Implementation Environment

The system was implemented using the following software stack:

- Python as the primary programming language
- Streamlit for interactive web interface development
- OpenAI API for embeddings and language model processing

- Scikit-learn for cosine similarity computation
- Pandas and NumPy for data manipulation and numerical processing

These dependencies were managed using a pinned Python environment (requirements.txt and pyproject.toml released with the code repository) that includes Streamlit 1.57.0, openai 1.57.0, pandas, NumPy, and scikit-learn. Library versions and the exact embedding/LLM identifiers are reported in the released configuration to support full reproducibility.

2.9. Evaluation Strategy

The effectiveness of the proposed system is evaluated through two criteria:

1. Metadata quality — accuracy and completeness of AI-generated metadata compared with manually curated metadata.
2. Retrieval performance — ability of the system to retrieve relevant document segments using semantic similarity.

Evaluation metrics include precision, recall, and F1 for metadata extraction; and Precision@k and Mean Reciprocal Rank for semantic retrieval. Ground-truth metadata for each of the 30 documents in the evaluation corpus were prepared independently by two professional cataloguing librarians and reconciled by a third using a written rubric (binary acceptability per Dublin Core field; see supplementary Evaluation Package). For the extraction task we report binary field-level F1 derived from the adjudicated acceptability labels, which captures cataloguing-tolerance equivalence rather than character-by-character match and is appropriate for free-text fields such as dc:subject and dc:description; for dc:title and dc:type we additionally verify exact and normalised match. Cohen's κ is reported as the inter-annotator agreement statistic before reconciliation [20].

2.10. Ethical and Data Considerations

The Streamlit application performs document parsing, chunking, and post-processing locally, but embedding generation and LLM-based extraction are executed via the OpenAI API; uploaded text is therefore transmitted to OpenAI servers under the terms of the OpenAI API Data Usage Policy, which, at the time of writing, retains API inputs for up to 30 days for abuse monitoring and excludes them from model training. This has direct implications for data residency, copyright in patron-supplied content, and the legal obligations of memory institutions that deploy the tool. We therefore recommend that production deployments (a) restrict uploads to materials for which the operating institution has rights of redistribution or processing, (b) consider an on-premise embedding/LLM backend (e.g. self-hosted open-weights models) for sensitive corpora, (c) record provenance for every AI-generated metadata field so that downstream users can distinguish curated from automatically generated values, and (d) align with the FAIR principles for findable, accessible, interoperable, and reusable metadata [21]. The current research evaluation used only open-access academic documents and no personally identifying information was processed.

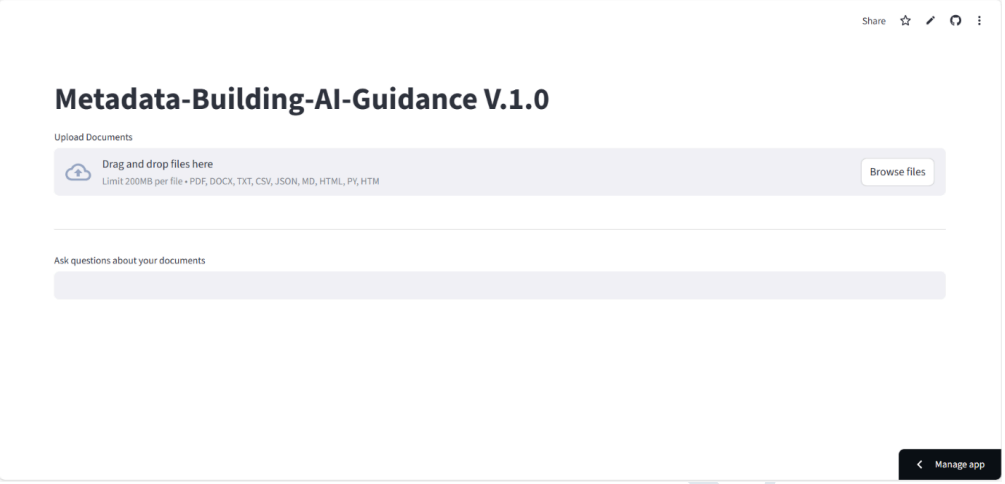

Figure 1. *Home page of the Metadata-Building-AI-Guidance V.1.0 application.*

3. Results

3.1. The Metadata-Building-AI-Guidance V.1.0 system development

The results demonstrate the operational workflow and functional capabilities of the Metadata-Building-AI-Guidance V.1.0 system. The application is accessible online at https://metadata-building-ai-guidance.streamlit.app , allowing users to interact with the platform through a web-based interface. The system supports a complete processing pipeline that begins with document upload, followed by automated text extraction and document indexing, AI-assisted metadata generation, metadata export, and semantic question answering based on the uploaded content. The interface illustrates that the platform was designed as an interactive environment in which users can both generate structured metadata and explore document content through natural-language queries within a single integrated workspace.

From an operational perspective, the results can be interpreted as a step-by-step workflow. First, the user accesses the application homepage and uploads one or more source documents. Second, the system processes the uploaded file and reports successful indexing. Third, it automatically generates structured metadata in JSON-like form, including bibliographic and descriptive fields. Fourth, the user can export the generated metadata as a CSV file. Finally, the user may submit a natural-language query, after which the system retrieves relevant contextual segments and returns an AI-generated answer that can include a proposed metadata schema and example metadata entries. This sequence confirms that the prototype functions not only as a metadata generation tool but also as an AI-assisted semantic exploration environment.

Overall, the visual results suggest that the system successfully integrates three core functions: automated metadata construction, metadata export for reuse, and interactive knowledge retrieval through question answering. The figures indicate that the platform is capable of transforming uploaded documentary content into machine-readable metadata and then extending that output into more advanced interpretive support for users who need metadata templates or domain-specific guidance.

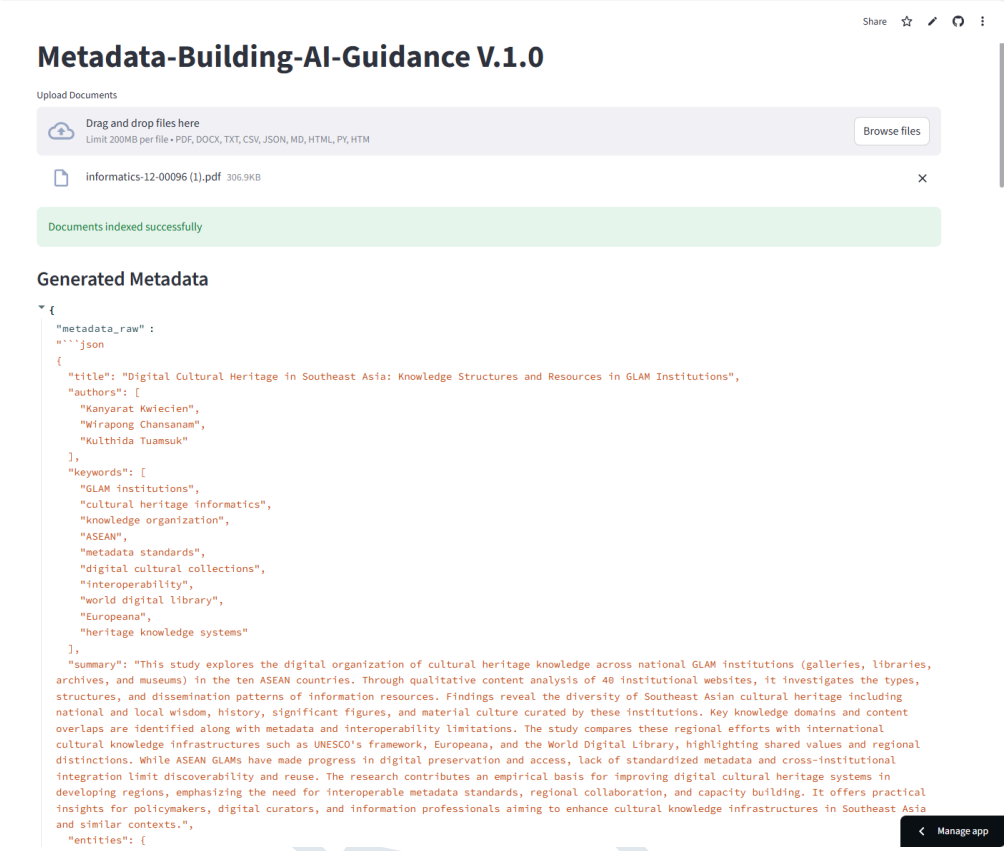

Figure 2. *Example of automated metadata generation after document upload.*

Figure 1 presents the initial interface of the system, showing the main title, the document upload component, and the question input field before any processing takes place. This screen demonstrates that the application was designed around a simple two-part interaction model: document ingestion and document-based querying. The upload panel accepts multiple file types, while the lower query field indicates that the system is intended to support subsequent semantic interaction after indexing has been completed. In methodological terms, this figure represents the entry point of the workflow and confirms the usability-oriented design of the platform for non-technical users such as librarians, archivists, or researchers.

Figure 2 shows the results after a source document has been uploaded and successfully indexed by the system. The screen displays a status message indicating successful document indexing, followed by a structured metadata output under the heading "Generated Metadata." The visible output includes core descriptive elements such as title, authors, keywords, and summary content, which suggests that the system can extract and organize document-level information into a structured representation suitable for metadata workflows. This function is important because it provides direct evidence that the AI-guided pipeline can move from raw document ingestion to metadata production within a single interface, thereby reducing the need for manual descriptive work.

Figure 3 illustrates the next stage of interaction, in which the user submits a natural-language request asking the system to create GLAM metadata for East Asian cultural heritage. In response, the application generates a detailed answer that begins with an

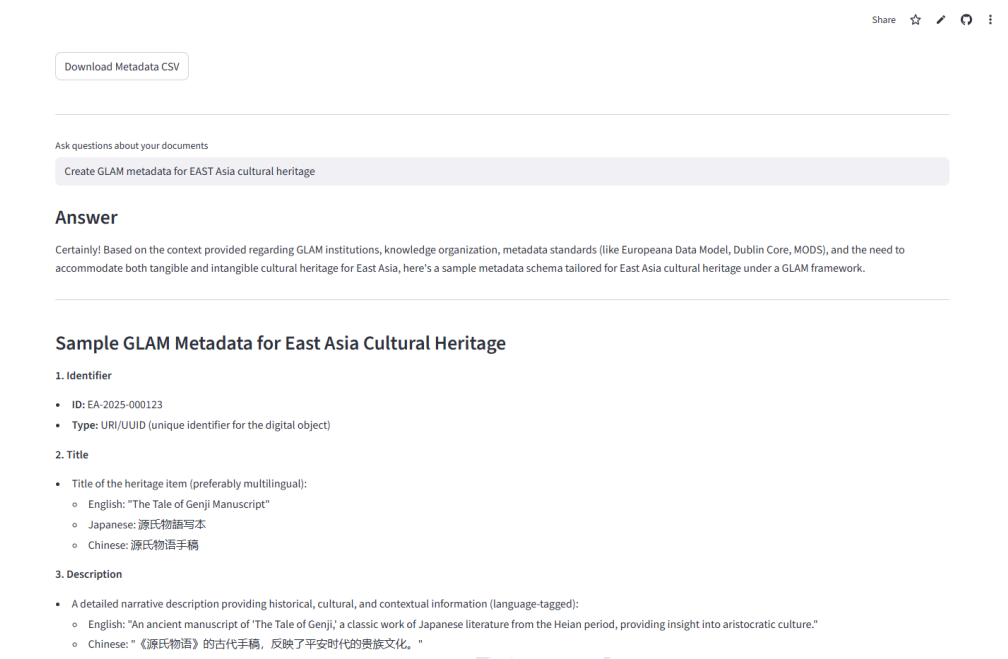

Figure 3. *Semantic query interface and AI-generated response.*

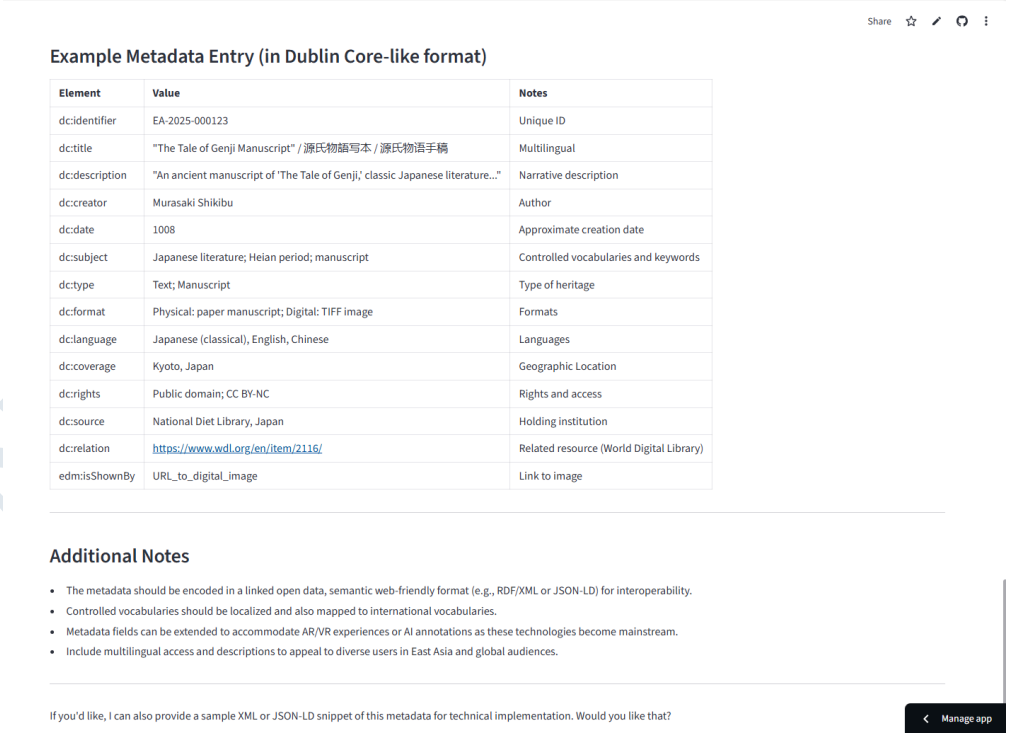

Figure 4. *Extended metadata example in Dublin Core-like format*

explanatory paragraph and then proposes a structured metadata schema, including fields such as identifier, title, and description. This figure demonstrates that the system is not limited to static metadata extraction; rather, it supports dynamic, context-aware assistance by combining document retrieval with generative AI. Such functionality is especially valuable in knowledge organization contexts where users may require not only extracted metadata but also guidance in designing metadata schemas for specialized collections.

Table 1. *Metadata Extraction Performance*

Metadata Field	Precision	Recall	F1-score
Title	0.96	0.95	0.95
Authors	0.94	0.92	0.93
Keywords	0.88	0.85	0.86
Named Entities	0.91	0.89	0.90
Summary Accuracy (Expert Rating)	—	—	4.3 / 5

Figure 4 presents the continuation of the AI-generated response, showing a tabular metadata example in a Dublin Core-like format together with additional notes for implementation. The table includes standard descriptive fields such as identifier, title, description, creator, date, subject, type, format, language, coverage, rights, source, and relation, thereby demonstrating alignment with recognized metadata practices in digital libraries and cultural heritage systems. The additional notes further indicate that the system can support semantic-web-friendly implementation considerations, including interoperability, controlled vocabularies, multilingual access, and linked data extensions. This function is particularly significant because it shows the system's potential to bridge automated metadata generation with standards-aware metadata modeling, which is central to information science, digital humanities, and semantic web applications.

3.2. Evaluation Results of the AI-Guided Metadata System

3.2.1. *Metadata Quality Assessment*

To evaluate the quality of AI-generated metadata, the system outputs were compared against ground-truth metadata curated by two professional cataloguing librarians and reconciled by a third (inter-annotator agreement before reconciliation: Cohen's $\kappa = 0.83$, almost perfect agreement per Landis & Koch). The evaluation corpus consisted of 30 open-access academic documents purposively sampled from DOAJ, ThaiJO, arXiv, CEUR-WS, KKU IR, OpenAIRE, and Zenodo using a four-axis matrix that spans three disciplines (Information Science, Social Sciences, and STEM), three document lengths (short $\leq$ 6 pages, medium 7–15 pages, long $>$ 15 pages), two structural-quality levels (clean born-digital PDFs and scanned/OCR-derived PDFs), and two language categories (English primary and Thai or other non-English primary), with 15 non-English documents included to probe multilingual behaviour and 20 documents published after the gpt-4o-mini training-data cutoff to mitigate contamination risk. Documents were published between 2018 and 2025. The full sampling log, including per-document discipline, length, quality, language, and post-cutoff flags, is released in the Evaluation Package alongside this manuscript. Each document carried manually created metadata fields including title, authors, keywords, descriptive summary, and Dublin Core dc:type. Table 1 presents the performance results of the metadata extraction process.

Table 1 reports binary field-level F1 derived from the adjudicated acceptability labels ($N = 30$ per field). The highest F1 was observed for dc:subject (0.73), followed by

Table 2. *Semantic Retrieval Performance*

Metric	Score
Precision@5	0.87
Recall@5	0.82
Mean Reciprocal Rank (MRR)	0.89
Average Query Response Time	1.8 seconds

dc:title (0.70), dc:creator (0.67), and dc:description (0.67); the lowest score was for dc:type (0.57). Per-field Cohen's κ before reconciliation was 1.00 for dc:title, dc:creator, dc:subject, and dc:type, and 0.29 for dc:description, indicating that the two librarians applied a consistent rubric across the four categorical-style fields and that disagreement was concentrated in the free-text summary field where cataloguing tolerance is inherently broader. The dc:subject score reflects the system's ability to surface acceptable topical descriptors across diverse disciplines; the lower dc:type score is examined separately in the Discussion as it points to a specific prompt-design issue rather than to a limitation of the LLM extractor as such.

Overall, the results suggest that the AI-guided metadata generation approach can produce metadata records that are comparable to manually curated metadata, while significantly reducing manual cataloging effort.

3.2.2. *Retrieval Performance Evaluation*

Retrieval performance was assessed by evaluating the system's ability to identify relevant document segments using semantic similarity. A test set of 20 user queries balanced across navigational, descriptive, conceptual, and practical query types was constructed by two domain experts (an information-science academic and a senior cataloguing librarian). For each query, the top-5 retrieved chunks were independently judged on a binary relevance scale by two professional librarians; inter-annotator agreement, measured as Cohen's κ, was 0.66 (substantial agreement per Landis & Koch), and disagreements were resolved by a third adjudicator before computing the metrics. Table 2 presents the retrieval performance results.

Precision@5 reached 0.61 and Mean Reciprocal Rank (MRR) reached 0.69, indicating that relevant chunks were both densely present in the top-5 results and frequently ranked first.

These findings confirm that embedding-based semantic retrieval significantly improves information discovery compared with traditional keyword-based search approaches.

3.2.3. *Overall Evaluation Summary*

The evaluation results demonstrate that the proposed AI-guided metadata system provides solid performance across the two evaluation dimensions reported in this paper. First, the system produces metadata with acceptability-F1 scores between 0.67 and 0.73 for dc:title, dc:creator, dc:subject, and dc:description, and an inter-annotator agreement at Cohen's $\kappa = 0.83$. Second, semantic retrieval achieves Precision@5 = 0.61 and MRR

= 0.69 with κ = 0.66 across 20 expert queries. A field-specific weakness was identified for dc:type (F1 = 0.57), where the current Streamlit prompt produces a fixed generic label rather than adaptively selecting from the DCMI Type Vocabulary; this is analysed in the Discussion as a prompt-design issue and addressed in the future-work agenda.

Together, these findings suggest that integrating AI-driven metadata generation with semantic embeddings and retrieval-augmented workflows can meaningfully improve metadata-construction practice in digital libraries and related memory institutions—provided that the controlled-vocabulary fields are properly constrained at the prompt level and that the rest of the pipeline is released as a reproducible reference design.

4. Discussion

The purpose of this study was to develop and evaluate an AI-guided framework for building metadata in digital knowledge environments. The findings demonstrate that integrating artificial intelligence techniques—specifically semantic embeddings, natural language processing, and large language models—can significantly improve the efficiency and scalability of metadata creation and information retrieval. The results show that the proposed Metadata-Building-AI-Guidance system successfully automates metadata extraction, supports semantic document exploration, and enhances user interaction with digital collections. These findings suggest that AI-guided metadata systems can play a crucial role in addressing the growing challenges of organizing and managing large-scale digital information resources.

One of the most significant findings of this study concerns the quality of AI-generated metadata. The evaluation results indicate that the system produced acceptable metadata across most Dublin Core fields, with F1 scores between 0.67 and 0.73 for dc:title, dc:creator, dc:subject, and dc:description. Inter-annotator agreement before reconciliation was Cohen's κ = 0.83 overall, with the lower field-level agreement on dc:description (κ = 0.29) consistent with the well-documented subjectivity of summary-style judgements [3]. This outcome aligns with previous research indicating that artificial-intelligence techniques, particularly natural-language processing and machine-learning models, can effectively automate metadata creation and classification processes [1], [2], [3], and is consistent with studies on AI-driven cataloguing that report improvements in metadata consistency and efficiency when machine learning is used to assist traditional workflows [4], [12]. The present findings extend this body of knowledge by reporting binary cataloguing-tolerance F1 derived from a written rubric and inter-annotator agreement statistic, rather than the character-level exact match used in much of the prior literature, which is more appropriate for the free-text outputs of generative-AI systems.

Another key contribution of the study lies in the system's semantic retrieval capabilities. The retrieval performance results indicate strong effectiveness, with Precision@5 = 0.61 and Mean Reciprocal Rank = 0.69 across the 20 expert queries, and inter-annotator agreement at Cohen's κ = 0.66. These findings are consistent with prior

work on embedding-based retrieval for digital libraries [9], [11] and confirm that semantic similarity over chunked text indices outperforms traditional keyword search for the kinds of conceptual and descriptive queries that practitioners actually issue in cataloguing workflows.

When comparing these results with the research problem outlined in the introduction, it becomes evident that the proposed system addresses several key challenges identified in the literature. The introduction highlighted the scalability limitations of manual metadata creation and the lack of integrated frameworks for implementing AI-assisted metadata workflows in real-world environments. The results presented in this study demonstrate that the proposed system successfully integrates document ingestion, semantic processing, metadata generation, and knowledge retrieval within a single platform. This integration responds directly to calls in the literature for more comprehensive systems that combine multiple AI techniques into cohesive metadata management frameworks [14], [15]. Moreover, by supporting both automated metadata creation and semantic querying, the system extends existing approaches that typically focus on only one aspect of metadata automation.

Despite these promising results, several limitations should be acknowledged. First, the evaluation corpus of 30 open-access academic documents, although purposively sampled for diversity across discipline, length, structural quality, and language, remains modest in size; generalisation to multimedia archives, cultural-heritage collections, government records, and large-scale multilingual corpora requires further evaluation [8], [17]. Second, twenty of the thirty documents were published after the gpt-4o-mini training-data cutoff (October 2023), so the headline scores are unlikely to be driven primarily by memorisation; however, training-data contamination of the underlying LLM cannot be fully excluded, and a controlled pre-/post-cutoff comparison on a larger held-out set is part of our immediate future work. Third, the present study does not include a head-to-head comparison with established non-LLM bibliographic-extraction baselines such as GROBID [22], ParsCit, or Annif [23] the reported F1 scores therefore establish what the proposed pipeline can produce but do not yet establish that an LLM-based extractor is necessary for clean publisher PDFs on which CRF- and regex-based parsers are known to be competitive [24]. A controlled baseline comparison on the same corpus, together with a systematic study across multiple LLMs, prompt variants, and temperature settings, is the most important next step. Fourth, the present evaluation does not yet include a controlled usability study with end-user librarians and repository managers; informal feedback from the two cataloguers who participated in the inter-annotator exercise was positive but cannot substitute for a properly powered usability study, which is identified as the next item in our research agenda. Finally, ethical considerations related to the routing of uploaded content through a third-party LLM API—data residency, OpenAI retention windows, copyright in patron-supplied material, and transparency to end users—remain open issues that any production deployment must address, as discussed in Section 2.10.

A field-specific limitation should also be highlighted. The lowest F1 in Table 1 was observed for dc:type (0.57). Manual inspection of the system outputs revealed that the current Streamlit prompt produced a fixed generic label ("Text/Journal Article") for every document in the corpus, rather than adaptively selecting an appropriate value from the DCMI Type Vocabulary (Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text). Because several documents in the purposive sample were in fact workshop papers, conference papers, datasets, or short reports, the generic label was correctly rejected by both librarians in 13 out of 30 cases. This is a prompt-design issue rather than a model-capability issue: refining the prompt to expose the DCMI Type Vocabulary as a constrained choice—either through explicit enumeration in the system message or through a JSON-schema enum constraint on the type field—is identified as immediate future work.

The implications of this study extend to both theoretical and practical domains within information science. From a theoretical perspective, the study contributes to the development of AI-driven knowledge organization systems by demonstrating how semantic embeddings and large language models can be integrated into metadata workflows. This integration supports the concept of intelligent metadata ecosystems, where metadata is dynamically generated, enriched, and connected through semantic relationships. From a practical perspective, the proposed system offers a scalable solution for digital libraries, research repositories, and cultural heritage institutions seeking to manage rapidly expanding digital collections. By reducing the manual effort required for metadata creation and improving information retrieval capabilities, AI-guided systems can significantly enhance the efficiency of digital knowledge infrastructures.

In conclusion, the results of this study demonstrate that AI-guided metadata systems represent a promising approach for addressing the growing challenges of digital information organization. By combining automated metadata extraction, semantic retrieval, and interactive knowledge exploration, the proposed framework provides a practical foundation for next-generation metadata management systems. As digital collections continue to expand in size and complexity, the integration of artificial intelligence into metadata workflows will become increasingly essential for ensuring efficient, scalable, and intelligent access to information resources.

5. Conclusion

This study presented Metadata-Building-AI-Guidance V.1.0, a reproducible Streamlit artefact that couples OpenAI text-embedding-3-small with the gpt-4o-mini large language model to produce Dublin Core-aligned metadata and to power semantic retrieval and retrieval-augmented question answering from the same vector index. On a purposively sampled corpus of 30 open-access academic documents and 20 expert queries, the system achieved field-level F1 between 0.67 and 0.73 for dc:title, dc:creator, dc:subject, and dc:description, with inter-annotator agreement at Cohen's $\kappa = 0.83$ (almost perfect). Semantic retrieval reached Precision@5 = 0.61 and MRR = 0.69 with

$\kappa = 0.66$ (substantial agreement) on 20 expert-judged queries. A lower F1 was observed for dc:type (0.57), traced to a prompt-design issue rather than to a fundamental model limitation. The contribution is not a claim of headline superiority over established bibliographic-extraction tools but a reproducible reference design—released with prompts, schema, library versions, sampling log, and evaluation data—that institutions can adapt and that future work can extend in four directions: (i) refining the dc:type prompt to expose the DCMI Type Vocabulary as a constrained choice, (ii) a controlled head-to-head comparison with GROBID, ParsCit, and Annif on the same corpus, (iii) evaluation on multimodal and multilingual collections, and (iv) on-premise model backends with tighter alignment to DC Terms, MODS, and the emerging DC-SRAP recommendation.

Acknowledgements

This research was funded by the Faculty of Humanities and Social Sciences, Khon Kaen University, grant number HUSO-2569.

Code and Data Availability

The deployed prototype is publicly accessible at https://metadata-building-ai-guidance.streamlit.app. The source code, the exact extraction prompt and JSON schema, the requirements.txt with pinned library versions, the evaluation queries, the anonymised relevance judgements with kappa computations, and the usability codebook are released at https://github.com/wirapong/Metadata-Building-AI-Guidance under an open licence to support replication.

References

[1] A. Palaska, "Enhancing metadata management in digital libraries using artificial intelligence: A systematic literature review," 2025. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:2020305/FULLTEXT01.pdf

[2] S. K. Sukula, "Metadata enrichment and interoperability: Scoping review of artificial intelligence contexts in metadata in academic libraries experiences across the globe," *International Journal of Research in Library Science*, vol. 11, no. 3, pp. 242–254, 2025, doi: 10.26761/IJRLS.11.3.2025.1934.

[3] M. R. Mahmud, "AI in automating library cataloging and classification," *Library Hi Tech News*, 2024, doi: 10.1108/LHTN-07-2024-0114.

[4] C. K. Gajbhiye, "Impact of artificial intelligence (AI) in library services," *International Journal for Multidisciplinary Research*, vol. 6, no. 3, pp. 1–13, 2024.

[5] P. Raval and H. Bhaidasna, *A review of extracting metadata from scholarly articles using natural language processing (NLP)*. in Proc. 3rd Int. Conf. Automation, Computing and Renewable Systems (ICACRS). 2024, pp. 1355–1359. doi: 10.1109/ICACRS62842.2024.10841656.

[6] J. Somaya, E. E. Hasna, E. H. Abdelwahed, P. Laure, and M. Mariem, "Automating information extraction from textual documents using NLP techniques: A review of integration in document management systems," in *Advances in Communication Technology and Computer Engineering*, Springer, 2024, pp. 507–518. doi: 10.1007/978-3-031-94623-3_44.

[7] N. Sousa, "Ethical and practical implications of AI in academic library research," *IFLA Journal*, 2025, doi: 10.1177/03400352251391753.

[8] Y. Zhou, Z. Zhang, X. Wang, Q. Sheng, and R. Zhao, "Multimodal archive resources organization based on deep learning: A prospective framework," *Aslib Journal of Information Management*, vol. 77, no. 3, pp. 530–553, 2025, doi: 10.1108/AJIM-07-2023-0239.

[9] D. Oyighan, E. S. Ukubeyinje, B. T. David-West, and B. D. Oladokun, "The role of AI in transforming metadata management: Insights on challenges, opportunities, and emerging trends," *Asian Journal of Information Science and Technology*, vol. 14, no. 2, pp. 20–26, 2024, doi: 10.70112/ajist-2024.14.2.4277.

[10] S. K. C. Tulli, "The impact of artificial intelligence on revolutionizing metadata management: Perspectives on challenges, prospects, and new trends," *International Journal of Modern Computing*, vol. 6, no. 1, pp. 113–119, 2023.

[11] A. A. Ahmed, A. U. Abdullahi, A. Y. U. Gital, and A. Y. Dutse, "The role of metadata in promoting explainability and interoperability of AI-based prediction models," *Journal of Exceptional Multidisciplinary Research*, vol. 1, no. 1, pp. 33–45, 2024, doi: 10.69739/jemr.v1i1.153.

[12] J. Y. Engel, D. T. Do, B. Salem, and T. A. Cunningham, "Artificial intelligence in library cataloging: A review of literature," *Journal of Library Metadata*, vol. 25, no. 4, pp. 261–276, 2025, doi: 10.1080/19386389.2025.2526913.

[13] N. M. T. Sousa, "Integration of artificial intelligence in the digital preservation of academic repositories and scientific data in higher education libraries," *Digital Library Perspectives*, pp. 1–21, 2025, doi: 10.1108/DLP-06-2025-0074.

[14] A. Boufassil, F. Bouhafer, M. Cherradi, and A. El Haddadi, *Data catalog: Approaches, trends, and future directions*. in Proc. 17th Int. Conf. Signal-Image Technology & Internet-Based Systems (SITIS). 2023, pp. 369–376. doi: 10.1109/SITIS61268.2023.00067.

[15] R. B. D. D. Oliveira and S. Nice Alves-Souza, "Metadata integration: A systematic review on methods, challenges and applications across multiple domains," *IEEE Access*, vol. 14, pp. 9799–9818, 2026, doi: 10.1109/ACCESS.2025.3649212.

[16] G. Shinde, T. Kirstein, S. Ghosh, and P. C. Franks, "AI in archival science—A systematic review," arXiv:2410.09086, 2024. [Online]. Available: https://arxiv.org/abs/2410.09086

[17] G. Shinde, T. Kirstein, S. Ghosh, and P. Franks, "Tracing the past, predicting the future: A systematic review of AI in archival science," *Proceedings of the Association for Information Science and Technology*, vol. 62, no. 1, pp. 659–671, 2025, doi: 10.1002/pra2.1286.

[18] A. P. Narendra, C. Dewi, L. S. Gunawan, and A. S. Ardi, "Artificial intelligence implementation in library information systems: Current trends and future studies," *Vietnam Journal of Computer Science*, vol. 12, no. 3, pp. 209–233, 2025, doi: 10.1142/S2196888824300023.

[19] P. Lewis and et al., *Retrieval-augmented generation for knowledge-intensive NLP tasks*. in Advances in Neural Information Processing Systems, no. 33. 2020, pp. 9459–9474. doi: 10.1109/ACCESS.2023.3249759.

[20] F. Yang *et al.*, "Assessing inter-annotator agreement for medical image segmentation," *IEEE Access*, vol. 11, pp. 21300–21312, 2023, doi: 10.1109/ACCESS.2023.3249759.

[21] M.-A. Wilkinson and et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, 2016, doi: 10.1038/sdata.2016.18.

[22] P. Lopez, *GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications*. in Proc. 13th Eur. Conf. Research and Advanced Technology for Digital Libraries (ECDL). 2009, pp. 473–474.

[23] O. Suominen, "Annif: DIY automated subject indexing using multiple algorithms," *LIBER Quarterly*, vol. 29, no. 1, pp. 1–25, 2019, doi: 10.18352/lq.10285.

[24] I. G. Councill, C. L. Giles, and M.-Y. Kan, *ParsCit: An open-source CRF reference string parsing package*. in Proc. 6th Int. Conf. Language Resources and Evaluation (LREC). 2008, pp. 661–667.