DCPAPERS

DCMI 2026

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

ISSN: 1939-1366

Grounding AI Subject Cataloguing in Standards and Policy: An MCP Server for Live LC Authority Lookup and a DITA-Encoded SHM for RAG

May Chan [1]

[1] University of Toronto

Abstract

Large language models (LLMs) applied to subject cataloguing using Library of Congress Subject Headings (LCSH) tend to generate headings and strings that are syntactically plausible but policy-invalid, bypassing the controlled vocabularies and governing rules that subject collocation depends on. This paper describes a two-track experiment to address the lack of grounding in standards and policy. The first track is `lc-vocabularies-mcp`, a Model Context Protocol (MCP) server connecting an LLM to live Library of Congress (LC) linked data APIs, enabling real-time authority validation for LCSH and related controlled vocabularies. The second track is a conversion of the *Subject Headings Manual* (SHM) from PDF to structured DITA (Darwin Information Typing Architecture), designed as a machine-actionable retrieval-augmented generation (RAG) corpus. Together, the two tracks support a five-part subject cataloguing workflow in which non-parametric knowledge is supplied to Claude at each part where parametric knowledge alone is insufficient. The paper reports on the architecture of `lc-vocabularies-mcp`, the DITA conversion methodology, and an evaluation design that tests whether policy-grounded retrieval improves AI-assisted subject cataloguing.

1. Introduction

Artificial intelligence (AI) is increasingly proposed and applied as a tool for subject cataloguing, a standards-governed professional practice in which controlled vocabularies determine what headings are valid and policy documentation determines how they are formulated and assigned. In practice, however, AI systems tend to bypass these standards [1], [2], relying on training data rather than authoritative sources. Outputs based on training data can resemble folksonomy, the improvised assignment of tags without reference to a controlled vocabulary or established standards [3]. Uncontrolled tagging circumvents the referential and relational semantic structures that subject vocabularies use to support subject collocation. This circumvention undermines the interoperability and consistency that make shared cataloguing possible [4], a risk that AI output compounds by being prolific and, without expert review, difficult to distinguish from valid headings.

Library of Congress Subject Headings (LCSH) are not simple tags: they are authorized vocabulary terms whose assignment and assembly into pre-coordinated strings must conform to a complex system of rules. An LCSH string such as *Architecture – Ontario – Toronto – History – 20th century* requires retrieving components from a controlled vocabulary and constructing the complete string according to rules codified in the *Subject Headings Manual* (SHM) [5]. The SHM is a Library of Congress (LC) policy

Full Paper

Available Aug 01, 2026

Correspondence to
May Chan
ms.chan@utoronto.ca

document of over 300 instruction sheets covering heading assignment, subdivision application, string assembly, and new heading proposals.

Large language models (LLMs) have limited access to this policy layer. Without it, they tend to produce headings that are syntactically plausible but policy-invalid [6]. Scenarios include unauthorized terms assigned, headings assembled in incorrect order, and free-floating subdivisions used outside their authorized contexts [1], [2]. My preliminary tests with `lc-vocabularies-mcp`, the MCP server described in Section 3, showed that authority file lookups alone were insufficient to ensure defensible output for every stage of the subject cataloguing workflow. When asked to justify its string construction and certain heading assignment decisions, Claude (Anthropic's LLM) acknowledged drawing on training rather than citing policy documentation. In my experience as a practitioner, human non-specialists tend to produce the same kinds of errors, reflecting insufficient awareness of, or regard for, policy requirements and reinforcing a broader undervaluation of cataloguing as a standards-governed practice.

A deeper structural reason may underlie the policy-invalid output of LLMs. Cataloguing tends to be characterized in publicly available text as data entry or keyword assignment rather than as the professional practice it actually is. This shallow characterization becomes the signal LLMs learn from, with consequences for the interoperability and consistency on which shared cataloguing depends. LLMs operating on such text function as pattern-matching systems, learning statistical relationships between tokens (words or word fragments) rather than acquiring the logical reasoning and domain knowledge that cataloguing policy requires [7]. This skewed distribution of shallow versus expert-level descriptions may compound the pattern-matching tendency, as models learn the surface form of headings without learning the policy logic that governs their construction [8]. The SHM codifies that policy logic in precise, professional language, but its PDF-only distribution means that even if it appears in training data, it does so as flat unstructured text that lacks the machine-actionable structure needed for reliable policy reasoning at inference time (when the model is actively generating a response).

Existing work on automated LCSH assignment has demonstrated the value of semantic similarity in suggesting base headings [9], [10]. However, this work does not address the policy layer governing how headings are validly selected and assembled, a challenge that requires a different approach. This paper describes a two-track experiment and outlines an evaluation designed to test whether grounding in authoritative sources and policy documentation improves the validity of AI-assisted subject cataloguing.

2. Decomposing the Subject Cataloguing Workflow

The two tracks are designed to support a five-part AI-assisted subject cataloguing workflow: subject analysis and candidate generation, heading identification, heading selection, string construction, and heading assignment. At each sub-task, non-parametric knowledge is supplied to Claude when parametric knowledge alone is insufficient. Parametric knowledge is encoded in the LLM's weights, the numerical values adjusted

during training that determine how the model responds to input. Non-parametric knowledge is retrieved from external sources at inference time and supplied to the model's context window (the active session text the model can see). Throughout this workflow, parametric and non-parametric knowledge are in constant interplay.

- **Subject analysis and candidate generation** is the process of determining what a work is about from its content (e.g., reading an abstract, title, or description) and then proposing candidate subject terms that have not yet been verified against the LC authority files. Expert-authored cataloguing instructions supplied at session initialization in the AI system provide structured guidance on subject analysis across LC controlled vocabularies, primarily LCSH, but also the LC Name Authority File (LCNAF) and the LC Genre and Form Terms (LCGFT). Claude's parametric knowledge allows it to interpret and apply these instructions to bibliographic input.
- **Heading identification** involves querying the LC authority files to determine whether candidate subject terms correspond to authorized headings. Claude invokes `lc-vocabularies-mcp` (described in Section 3) on demand to supply non-parametric knowledge via API lookup [11], [12]. The server returns authorized labels, URIs, scope notes, geographic subdivision eligibility, collection membership (indicating whether a heading is a subdivision or pattern heading), full authority record data including variant labels and, for name authority records, biographical information.
- **Heading selection** involves mapping candidate subject terms to authorized headings by reasoning over the authority data already in context from heading identification. Claude applies its parametric knowledge to compare candidates, interpret scope notes, assess geographic subdivision eligibility, and use collection membership to identify headings that require special string construction treatment. It then selects the authorized heading that most accurately represents a concept identified during subject analysis.
- **String construction** is the assembly of authorized headings into valid pre-coordinated subject heading strings according to the SHM. Strings are constructed when the subject matter of the work cannot be adequately represented by a base heading. Rules governing subdivision order, geographic subdivision application, pattern heading consultation, and free-floating subdivision eligibility are not derivable from the authority file itself and cannot be reliably encoded in parametric knowledge alone. For example, determining what subdivisions may be added to a heading such as *Painting, Canadian* requires consulting the authorized pattern for art headings (SHM H 1148). While a free-floating subdivision such as *–History* carries collection membership data indicating applicable pattern heading categories, the detailed conditions governing its use under specific heading types are codified in the SHM. Claude would need to invoke the SHM MCP server (described in Section 4) on demand to retrieve the relevant policy guidance as non-parametric knowledge via

retrieval-augmented generation (RAG), applying its parametric knowledge to interpret and act on the returned rules. The retrieval mechanisms for the SHM instruction sheets are described in Section 4.2.

- **Heading assignment** is the act of assigning valid headings to the work being described, drawing on the context accumulated across prior sub-tasks. Claude applies its parametric knowledge to reason from this context to make the final assignments, which must be defensible with reference to any rules consulted during the workflow. In professional cataloguing practice, heading decisions should be grounded in the SHM, with cataloguers being able to identify the relevant instruction sheet and section that informed each decision. In this experiment, Claude is required to cite that rationale explicitly, making the reasoning transparent and auditable.

  Treating cataloguing as a single task risks applying AI inference when a lookup would be more reliable, or relying on lookup alone when policy reasoning is also required. Decomposing the workflow makes it possible to supply each sub-task with the most appropriate source of non-parametric knowledge. The evaluation described in Section 5 is designed to measure the validity of heading assignments produced by this approach.

3. Track One: `lc-vocabularies-mcp`

The first track of the experiment is `lc-vocabularies-mcp`, an MCP (Model Context Protocol) server that connects an LLM (in this case Claude, running in Claude Desktop under an institutional Claude Edu licence) to the LC Linked Data Service APIs. MCP is an open protocol developed by Anthropic through which AI models connect to external applications to retrieve non-parametric knowledge [13] on demand. This first track adopts a validation workflow proposed by Tang and Jiang, in which natural language candidate terms are queried against the LC Linked Data Service and validation results are returned to the LLM for refinement. Tang and Jiang implemented this workflow through three deployment approaches, including a proof-of-concept MCP server, `cataloger-mcp` [14]. Building on `cataloger-mcp`'s foundation, `lc-vocabularies-mcp` extends the approach to include all LCNAF name types and LCGFT, and adds richer authority data retrieval to support heading selection reasoning. This data includes scope notes, full authority record data, geographic subdivision eligibility, and collection membership. Claude may invoke these tools multiple times within a single cataloguing session, autonomously refining its heading candidates as each lookup informs the next decision. User intervention remains possible at any point, whereas in `cataloger-mcp` each refinement requires user initiation.

Openly available on GitHub [15], `lc-vocabularies-mcp` currently offers 16 tools organized around LCSH, LCNAF, and LCGFT. LCSH tools support left-anchored string and keyword search against the id.loc.gov `suggest2` API [16], scope note retrieval via the SKOS JSON endpoint, and full authority record retrieval via MADS (Metadata Authority Description Schema) RDF/XML parsing. LCNAF search tools cover personal

names, corporate names, geographic names, and meeting names, each queryable by left-anchored or keyword search. For LCGFT, the same `suggest2` infrastructure supports genre and form term lookup.

A key design decision is the automated enrichment of LCSH search results with geographic subdivision eligibility. For the first six most relevant results of every LCSH search, the server fetches the MADS/RDF JSON-LD record from named MADS collections at id.loc.gov and checks whether the heading is a member of the `collection_SubdivideGeographically` collection, returning a `maySubdivideGeographically` boolean field indicating whether geographic subdivision is permitted. This automated enrichment directly addresses a common string construction error among non-expert cataloguers, namely the application of geographic subdivisions to headings that do not permit them, an error that AI systems generating LCSH are also prone to. The cap of six results reflects SHM H 180′s general maximum of six subject headings per work and keeps API response times practical. This cap is currently enforced through cataloguing instructions supplied at session initialization, a temporary measure until the SHM MCP server described in the next section makes H 180 retrievable directly from the policy corpus at inference time.

The same MADS/RDF fetch also returns a `collectionMembership` field containing human-readable labels for any other MADS collections the heading belongs to, for example, “LCSH Collection — Subdivisions” or “Pattern Heading — Ethnic Groups”. A non-empty value indicates to Claude that the heading is a subdivision or pattern heading rather than a standard topical main heading, signalling that special string construction rules apply. The `collectionMembership` field also surfaces collection URIs that identify the relevant SHM instruction sheet for pattern headings. For example, the URI of the form `collection_PatternHeadingH1095` directly encodes the H-number, creating a deterministic link from authority lookup results to SHM H 1095 that the SHM MCP server could leverage to retrieve the relevant policy sheet.

The scope note retrieval tool fetches usage guidance from the SKOS JSON endpoint, supporting disambiguation when multiple authorized candidate headings appear similar. The full authority record tool parses RDF/XML to extract preferred labels, variant labels, LC classification numbers, external identifiers (VIAF, ISNI, Wikidata), administrative history, and biographical data (for name authority records). Together, these tools provide the additional context needed to select the correct heading. The data these tools return supports authority control decisions such as disambiguating personal names or confirming the scope of a topical heading.

4. Track Two: SHM Conversion to DITA for RAG

4.1. DITA Schema and Conversion Methodology

The second track addresses the lack of machine-actionable policy documentation. The SHM, publicly distributed as PDFs, contains the rules needed to validly construct and assign headings. Making these rules available to an AI system requires converting them

into a form that can be retrieved semantically, chunked at appropriate boundaries, and filtered by the type of guidance a cataloguer working with LC vocabularies would consult. The DITA corpus is designed as a retrieval layer that pairs expert-curated metadata filters with dense semantic search (explained in Section 4.2) to supply accurate policy reasoning at inference time [11], grounded in professional knowledge at the specificity and precision that subject cataloguing requires.

I have chosen DITA (Darwin Information Typing Architecture) [17] as the conversion format for its content typing discipline rather than its publishing capabilities. DITA's concept/task/reference distinction maps directly onto the different types of guidance SHM sheets contain: background context (concept), procedural rules (task), and lookup tables and lists (reference). A single instruction sheet will therefore produce multiple DITA topic files reflecting its internal structure; a sheet with distinct background, procedural, and lookup content will produce a corresponding set of concept, task, and reference files. Research in scholarly publishing has shown that XML's explicit element tagging, hierarchy, and modularity reduce ambiguity in RAG retrieval and allow systems to pull relevant content without processing redundant information [18].

The conversion project is establishing a DITA metadata schema specifically designed to enable filtered retrieval. Metadata is applied at two levels: each file carries metadata identifying the instruction sheet and the broad types of guidance it governs, while each chunk carries metadata encoding the specific type of guidance it contains, with values assigned through expert analysis of the source PDF rather than automatic text processing. This expert-curated metadata is what the hybrid retrieval mechanism is designed to depend on: filters will narrow the candidate pool to chunks relevant to the current cataloguing sub-task before vector similarity search ranks them by semantic similarity to the query. Embeddings will make the structured content semantically accessible, but the metadata will determine what is retrievable in the first place.

A critical design decision is the treatment of chunk boundaries in the instruction sheets, since each chunk will carry metadata classifying the type of guidance it contains. Automatic chunking by character count or sentence boundaries, the default approach in most RAG implementations, would split rules across chunks in ways that produce potentially incoherent retrievable units. Research has shown that semantic markup-based chunking, which identifies what content means (such as DITA's `<step>` or `<concept>` elements) rather than relying on surface formatting markers or token count, improves RAG accuracy and is more generalizable across document types [19]. The SHM DITA conversion applies this principle of semantic markup by bringing cataloguing expertise to chunk boundary decisions through a supervised pre-analysis worksheet completed before DITA conversion begins. Domain knowledge is applied at each stage, particularly for guidance type assignments and boundary placement. Accurate chunk boundaries are a precondition for accurate chunk-level metadata. A chunk spanning two different types of guidance cannot be described by a single metadata value, and metadata that does not accurately reflect chunk content will misdirect the retrieval system regardless

of embedding quality. Expert judgment in chunking therefore serves both retrieval mechanisms, ensuring semantically coherent boundaries and chunk-level metadata that faithfully reflects content.

Instruction sheets have been prioritized for conversion based on a tier structure. Tier 1 consists of core sheets governing general heading assignment and string construction, including H 80, H 180, and H 1075. Tier 2 adds sheets identified through cross-reference traversal from Tier 1 and manual selection based on my cataloguing experience in the areas of history, political science, and geography. Together, these sheets constitute the first meaningful RAG testing unit. The conversion process is itself an act of genre and content analysis: each sheet must be read, interpreted, and described at a level of granularity that makes its policy guidance retrievable and actionable, producing metadata that is, in effect, an indexed catalogue of the SHM.

4.2. SHM MCP Server and Retrieval Design

The DITA corpus will be exposed to Claude as the AI cataloguing agent via dedicated SHM MCP server instances currently under development. Where `lc-vocabularies-mcp` supplies non-parametric knowledge via deterministic API lookup, the SHM MCP server will supply it via hybrid retrieval: metadata filters first narrow the candidate pool to chunks matching the relevant heading type and guidance category, and vector similarity search then ranks the filtered candidates by semantic similarity to the retrieval query, which reflects the accumulated context of prior sub-tasks. This metadata filtering component draws on the principle of taxonomy-guided retrieval, in which domain-specific thematic structures filter the search space before similarity matching, with the DITA metadata schema functioning as an expert-curated cataloguing taxonomy [20]. The SHM MCP server will be invoked by Claude selectively to retrieve policy guidance for string construction and heading assignment as required. This selective invocation is consistent with Mallen et al.'s finding that retrieval augmentation is most valuable when LLMs cannot reliably encode knowledge from training data [21].[1]

The similarity component of the hybrid retrieval uses dense retrieval, in which documents and queries are encoded as vectors in a shared semantic space and similarity is measured by distance between them. A sentence transformer fine-tuned for semantic similarity will produce these embeddings. Dense retrieval is preferred over keyword-based approaches such as BM25 for this corpus on the grounds that cataloguing queries are unlikely to share exact terminology with the instruction sheet text they are seeking [22] semantic similarity is better suited to bridging that terminological gap than lexical matching. While the SHM corpus is English-only, the bibliographic inputs that an LLM reasons about, such as titles, abstracts, and subject queries, can be in languages other than English. A multilingual model is therefore preferable for handling linguistically diverse queries without degrading retrieval quality against the English-language policy corpus. The selected model is `paraphrase-multilingual-MiniLM-L12-v2` [23].

[1]For an overview of retrieval paradigms including sparse, dense, hybrid, and taxonomy-guided approaches, see [20].

Table 1. *Non-parametric knowledge supply and parametric knowledge roles in the subject cataloguing workflow*

Sub-task	Non-parametric knowledge source	Mechanism(s) involved	Role of Claude's parametric knowledge
Subject analysis and candidate generation	Project instructions; SKILL.md (session initialization)	Claude (decoder-only LLM)	Interprets and applies supplied instructions to bibliographic input
Heading identification	id.loc.gov (on demand)	`lc-vocabularies-mcp` MCP server (live API lookup)	Invokes the API lookup; determines when and what to query
Heading selection	Context from prior sub-tasks	Claude (decoder-only LLM)	Reasons over authority data already in context
String construction	SHM in DITA (on demand)	SHM MCP server (hybrid retrieval: metadata filtering + dense embeddings)	Invokes SHM retrieval; applies returned rules to assemble string
Heading assignment	Context from prior sub-tasks	Claude (decoder-only LLM)	Synthesizes prior outputs to justify decisions

The DITA files will be embedded, stored, and indexed in ChromaDB [24], a lightweight open-source vector database selected for its ease of local deployment and native support for metadata filtering alongside vector similarity search. The SHM MCP server will connect to ChromaDB to serve the hybrid retrieval requests. All processes will run on Arbutus Cloud, a research computing resource operated by the Digital Research Alliance of Canada.

5. Evaluation Design for the Integrated Workflow

The central hypothesis is that grounding an LLM in live authority data and machine-actionable policy documentation at inference time will produce more valid heading assignments than an ungrounded baseline. The two servers, `lc-vocabularies-mcp` and the SHM MCP server, will operate together within a single cataloguing session in Claude Desktop or any MCP-compatible AI tool. In the intended workflow, the AI cataloguing agent first invokes `lc-vocabularies-mcp` for authority lookup and then invokes the SHM MCP server for policy retrieval. Table 1 (overleaf) summarizes the knowledge sources, mechanisms, and the role of Claude's parametric knowledge across the full workflow.

Claude is a decoder-only LLM. The SHM MCP server retrieves policy guidance via hybrid retrieval, which combines metadata filtering with dense vector similarity search over SHM DITA corpus embeddings generated by a sentence transformer, in this case an encoder model fine-tuned for semantic similarity [9], [10]. For an overview of transformer architectures serving different functions in LLMs, see Liu [25].

The evaluation is designed around six comparators. The first is an ungrounded baseline in which Claude constructs and assigns headings solely from training knowledge. Existing literature documents poor LLM performance on subject heading assignment in general [1], [2]. The second comparator will add `lc-vocabularies-mcp` without any SHM retrieval, isolating the contribution of live authority lookup to heading validity.

The remaining four comparators will pair `lc-vocabularies-mcp` with SHM retrieval at increasing levels of structural sophistication, each assigned to a separate MCP server instance to prevent cross-collection context contamination during testing. All four ChromaDB comparators will use `paraphrase-multilingual-MiniLM-L12-v2` [23], the embedding model described in Section 4.2. The first two apply dense retrieval over PDF-derived collections, varying chunking strategy from full-sheet chunking to surface-marker-based chunking [7]. The third applies dense retrieval over the DITA corpus at file-level granularity, isolating the contribution of expert-determined chunk boundaries

Table 2. *Comparator design for the evaluation of AI-assisted subject cataloguing*

No.	SHM Corpus	Chunking	Retrieval Unit	Retrieval Method
1	—	—	—	Training knowledge only
2	—	—	—	Authority lookup only
3	Raw PDF	Full-sheet	One vector per sheet (312 total)	Authority lookup + dense
4	Raw PDF	Surface-marker-based[a]	One vector per section	Authority lookup + dense
5	DITA files	Expert-determined (DITA file-level)	One vector per DITA file	Authority lookup + dense
6	DITA files	Expert-determined (DITA file-level)	One vector per DITA file	Authority lookup + hybrid (metadata filtering + dense)

to retrieval quality. The fourth applies hybrid retrieval over the same DITA corpus, combining metadata filtering with dense vector similarity search, isolating the contribution of expert-curated metadata to policy grounding. The experiment is designed with the expectation that grounding quality will improve across this progression. Table 2 (overleaf) summarizes the corpus format, chunking strategy, retrieval unit, and retrieval method for all six comparators.

Comparators 2–6 all include `lc-vocabularies-mcp` for live authority lookup. Each SHM instruction sheet is converted into multiple DITA topic files; file-level retrieval granularity means one vector per topic file.

[a] Surface-marker-based chunking splits each PDF at recognizable section boundaries (e.g., the BACKGROUND section, numbered sections, and lettered sections) identified by surface text patterns rather than semantic markup [7].

Test cases will be drawn from cataloguing scenarios covering the sub-tasks outlined in Table 1, with pre-verified correct answers to be established in consultation with subject cataloguing experts. Validation will use prompt-level citation requirements in which Claude cites the SHM instruction sheet and section supporting each decision to make reasoning transparent and auditable. A structured feedback form is being developed to support evaluation of the first two comparators: the ungrounded baseline and heading identification and selection using `lc-vocabularies-mcp` only. Evaluation instruments for the RAG comparators will follow once their SHM MCP servers are operational.

6. Potential Directions and Implications

Improving the validity of outputs from AI-assisted subject cataloguing is not only a quality concern, but also an urgent one. The shallow characterization of cataloguing in publicly available text produces poor parametric knowledge in LLMs. Policy-invalid headings entering the bibliographic record degrade the shared catalogue and, if incorporated into future training data, compound both problems. Addressing these problems requires training data that accurately represent the standards-governed nature of cataloguing practice as well as tools grounded in authoritative sources.

Should the evaluation results support the hypothesis that policy-grounded retrieval improves heading validity, this experiment could be extended in several directions. The DITA conversion would be expanded to cover the full SHM corpus beyond the proof-of-

concept testing unit. The architecture could also be adapted to other cataloguing policy documentation such as the *Genre/Form Terms Manual* for LCGFT, and for LCNAF, the *NACO Participants' Manual* and *Descriptive Cataloging Manual* section Z1 (DCM Z1). The LC policy documentation on which this experiment depends would need to be monitored against updates, as LC transitions to LRM-aligned RDA and BIBFRAME. More speculatively, the structured outputs of this experiment could serve as training data for a purpose-built small language model (SLM) specialized for subject cataloguing [1], complementing the inference-time grounding approach with one that internalizes that grounding through training.

Two directions for extending the retrieval architecture are also worth pursuing. First, the evaluation described in this paper tests dense and hybrid retrieval; a broader comparison against keyword-based retrieval (BM25) and reranking approaches would establish whether RAG method choice materially affects retrieval quality. Second, the LC linked data APIs on which `lc-vocabularies-mcp` depends expose RDF graph structure that the current server does not fully exploit. Extending the server to traverse `skos:broader`, `skos:narrower`, and `skos:related` links, and to follow `owl:sameAs` connections to external knowledge graphs such as VIAF and Wikidata, would enable richer authority grounding than point lookup alone. Furthermore, the DITA `<xref>` element already encodes inter-sheet citation relationships as directed, machine-readable edges. A local metadata schema can extend these relationships further, capturing additional expert-determined relationships between sheets and chunks, such as functional groupings or concept-to-task dependencies, where the relationship is meaningful and consistent enough to be worth encoding. Whether such structures would support effective graph-based retrieval depends on relationship density across the corpus; the core instruction sheets are frequently co-cited, but density thins considerably across the larger corpus.

A complementary implication concerns linked open data infrastructure. Where controlled vocabularies are maintained as linked open data by standards bodies and public institutions, exposing that data through queryable APIs would make it directly accessible to AI agents, rather than only as bulk downloads or human-readable documentation. The relationship structure inherent in linked data offers additional reasoning potential beyond simple lookup, enabling AI agents to navigate between related entities. More widespread publication of linked open data could support more reliable AI-assisted work in any domain governed by established controlled vocabularies, as access to the LC linked data APIs via `lc-vocabularies-mcp` demonstrates.

The approach described in this paper carries implications for AI-assisted work in any domain in which policy governs practice; the gap between plausible and valid outputs is not unique to cataloguing [6]. Evidence from scholarly publishing and financial report processing suggests that documentation structured with RAG retrieval in mind (i.e., semantically tagged, chunked at meaningful boundaries, and filterable by guidance type) is a more effective grounding source for AI agents than unstructured prose [18],

[19]. The present work suggests that expert judgment in determining chunk boundaries adds value that automatic chunking cannot replicate.

7. Conclusion

Effective AI-assisted subject cataloguing requires more than parametric knowledge of a controlled vocabulary. It also requires grounding in live authority data and structured policy documentation, with non-parametric knowledge supplied to Claude at each subject cataloguing sub-task where parametric knowledge alone is insufficient. LLMs cannot reliably construct or assign valid subject headings without access to the SHM, which in its current PDF form lacks the machine-actionable structure needed for reliable reasoning. The two-track architecture described in this paper, `lc-vocabularies-mcp` for live authority lookup paired with an SHM MCP server for policy-grounded retrieval over a DITA-encoded corpus, is designed to address both limitations. The evaluation tests whether expert-determined chunk boundaries and metadata-enabled filtered retrieval improve policy grounding over automatic chunking approaches, and whether policy-grounded retrieval improves heading validity over authority lookup alone. These design decisions reflect a broader argument: that grounding AI systems in authoritative sources requires a principled account of what kind of knowledge each sub-task demands and why.

The conference theme of meaning-driven AI asks how metadata can align AI systems with human values. This project engages that question in the domain of professional practice. Cataloguing policy embodies accumulated expert judgment about how to represent knowledge in ways that are findable, consistent, and useful to researchers. Making that judgment machine-actionable is one way metadata can meaningfully shape AI behaviour. The DITA conversion aims to do so by enriching policy documentation with structured metadata that AI systems can reason from rather than bypass. When an AI system generates unauthorized terms or heading strings that violate the policy of a knowledge organization system it purports to serve, it is not merely making a technical error; it is eroding the integrity of that system, on which consistent application in a cooperative cataloguing ecosystem depends.

Chou and Chu call on cataloguing and metadata librarians to develop sufficient AI knowledge to collaborate effectively with AI and natural language processing experts to improve metadata quality [10]. This paper models that engagement, explaining how the components of an AI-assisted cataloguing system work and why each design decision was made. It also demonstrates that the domain expertise cataloguers bring to AI system design is not incidental but essential.

Acknowledgments

This work was made possible by a research leave approved by the University of Toronto Libraries (UTL), during which the ideas seeding the project were developed. The continued development of `lc-vocabularies-mcp` and the SHM DITA conversion has been supported by UTL's encouragement of ongoing experimentation with AI-assisted

cataloguing. The UTL AI Technology Readiness Working Group provided invaluable guidance on institutional access to the tools the project required.

Appendix A. AI Statement

This paper was developed with Claude (Anthropic, `claude-sonnet-4-6`), functioning as a collaborative assistant in drafting and revising the manuscript. The experiment described reflects my original research, domain knowledge, and professional expertise. Paper content was drawn from documentation produced in the course of the research and chat histories about MCP server design, schema development, and evaluation design. I take full responsibility for the accuracy of all claims and citations.

References

[1] C. Holstrom, "Large language models (LLMs) and cataloging: Exploring how ChatGPT and Copilot assign subject headings and call numbers," *NASKO*, vol. 10, 2025, doi: 10.7152/nasko.v7i1.95648.

[2] B. Dobreski and C. Hastings, "AI chatbots and subject cataloging: A performance test," *Library Resources & Technical Services*, vol. 69, no. 2, 2025, doi: 10.5860/lrts.69n2.8440.

[3] A. Mathes, "Folksonomies: Cooperative Classification and Communication Through Shared Metadata." [Online]. Available: https://adammathes.com/academic/computer-mediated-communication/folksonomies.html

[4] E. Svenonius, *The Intellectual Foundation of Information Organization*. MIT Press, 2000.

[5] Library of Congress, "Subject Headings Manual," 2008. [Online]. Available: https://www.loc.gov/aba/publications/FreeSHM/

[6] Malikussaid, H. H. Nuha, and I. Kurniawan, *Bridging the plausibility-validity gap by fine-tuning a reasoning-enhanced LLM for chemical synthesis and discovery*. in Proceedings of the 2025 IEEE 18th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC). 2025. doi: 10.1109/MCSoC67473.2025.00090.

[7] C. Davis, "Unlocking web archives: LLMs, RAG, and the future of digital preservation," University of Victoria Libraries, 2025. [Online]. Available: https://hdl.handle.net/1828/21379

[8] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, 2022, doi: 10.1145/3571730.

[9] N. Kazi, N. Lane, and I. Kahanda, *Automatically cataloging scholarly articles using Library of Congress Subject Headings*. in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-srw.7.

[10] C. Chou and T. Chu, "An analysis of BERT (NLP) for assisted subject indexing for Project Gutenberg," *Cataloging & Classification Quarterly*, vol. 60, no. 8, pp. 807–835, 2022, doi: 10.1080/01639374.2022.2138666.

[11] P. Lewis *et al.*, *Retrieval-augmented generation for knowledge-intensive NLP tasks*. in Advances in Neural Information Processing Systems, no. 33. 2020, pp. 9459–9474. doi: 10.48550/arXiv.2005.11401.

[12] Anthropic, "The Claude 3 model family: Opus, Sonnet, Haiku," Anthropic, 2024. [Online]. Available: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

[13] Anthropic, "Model Context Protocol." [Online]. Available: https://modelcontextprotocol.io/

[14] K.-L. Tang and Y. Jiang, "Better recommendations: Validating AI-generated subject terms through LOC Linked Data Service." [Online]. Available: https://doi.org/10.48550/arXiv.2508.00867

[15] M. Chan, "lc-vocabularies-mcp." [Online]. Available: https://github.com/msuicaut/lc-vocabularies-mcp

[16] Library of Congress, "Library of Congress Linked Data Service." [Online]. Available: https://id.loc.gov/

[17] OASIS DITA Technical Committee, "DITA Version 1.3 Specification," 2018. [Online]. Available: https://docs.oasis-open.org/dita/dita/v1.3/errata02/os/complete/part3-all-inclusive/dita-v1.3-errata02-os-part3-all-inclusive-complete.html

[18] M. Gross, *From valid XML to valuable XML: When "good" matters more than "valid"*. in Journal Article Tag Suite Conference (JATS-Con) Proceedings 2025. National Center for Biotechnology Information, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK611679/

[19] A. Jimeno Yepes, Y. You, J. Milczek, S. Laverde, and L. Li, "Financial report chunking for effective retrieval augmented generation." [Online]. Available: https://doi.org/10.48550/arXiv.2402.05131

[20] P. Jiang, S. Ouyang, Y. Jiao, M. Zhong, R. Tian, and J. Han, *A Survey on Retrieval and Structuring Augmented Generation with Large Language Models*. in Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2025. doi: 10.1145/3711896.3736557.

[21] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, *When not to trust language models: Investigating effectiveness of parametric and non-parametric memories*. in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546.

[22] V. Karpukhin *et al.*, *Dense Passage Retrieval for Open-Domain Question Answering*. in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

[23] N. Reimers, "paraphrase-multilingual-MiniLM-L12-v2." [Online]. Available: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[24] Chroma, "Chroma." [Online]. Available: https://github.com/chroma-core/chroma

[25] B. Liu, *Comparative Analysis of Encoder-Only, Decoder-Only, and Encoder-Decoder Language Models*. in Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024). SCITEPRESS, 2024. doi: 10.5220/0012829800004547.