

KG-LibRAG: A Conceptual Framework for Knowledge Graph-Driven Generative AI in Intelligent Information Retrieval and Personalised Resource Recommendation in Academic Libraries

Harihararao Mojjada

Assistant Librarian,

MVGR College of Engineering (A),

Vizianagaram, Andhra Pradesh, India, and

ORCID: <https://orcid.org/0000-0003-1737-6146>

Abstract:

Purpose – Discovery systems in academic libraries still rely largely on keyword matching over flat metadata, which limits semantic understanding, context-aware retrieval and personalisation, while generative AI offers conversational access but is prone to fabricated content. This paper proposes KG-LibRAG, a conceptual framework that grounds generative AI in a library knowledge graph for intelligent retrieval and explainable, personalised resource recommendation, and specifies how the framework should be evaluated.

Design/methodology/approach – The paper is conceptual. Following a design science approach, it derives design requirements from an integrative review of research on semantic retrieval, knowledge graphs, generative AI in libraries, hallucination and grounding, and scholarly recommendation, interpreted through a sociotechnical lens. These requirements inform the design of a five-layer architecture, five research propositions and a protocol for empirical evaluation.

Findings – The framework comprises five layers — library data sources, a library knowledge graph, a generative AI layer, hybrid retrieval and personalised recommendation — together with a cross-cutting governance function and a feedback loop. Its distinctive elements are an ontology that reuses BIBFRAME, SKOS and DCMI terms and adds course and consent-based user-profile classes; fusion of lexical, dense and graph retrieval; generation constrained to retrieved records, with every cited identifier verified before an answer is shown; and path-based explanations for recommendations. The paper states five testable propositions and an evaluation protocol comprising four comparison conditions, retrieval, recommendation, faithfulness and user-centred measures, sample-size targets and an analysis plan.

Research limitations/implications – The framework has not yet been implemented or empirically validated, and the review on which it rests is integrative rather than systematic. The propositions and protocol provide an agenda for testing it, beginning with a single-institution prototype.

Practical implications – The framework gives libraries running open-source systems such as Koha and DSpace a staged adoption path that does not require replacing their existing systems, and identifies new professional roles in knowledge engineering, AI curation and evaluation.

Originality/value – The paper unifies knowledge graphs, generative AI and personalisation within one architecture designed around library metadata standards and librarian oversight, with explicit safeguards against hallucinated citations and a ready-to-use protocol for evaluating it.

Keywords – Knowledge graphs, Generative AI, Large language models, Retrieval-augmented generation, Information retrieval, Recommender systems, Academic libraries, Sociotechnical systems, Design science, Conceptual framework

1. Introduction

1.1 Background

Academic libraries have spent three decades translating their collections into digital form. Online public access catalogues (OPACs), web-scale discovery layers, institutional repositories and consortial e-resource subscriptions now give users access to far more material than any print collection could hold. Yet the mechanism through which users find that material has changed surprisingly little. Most library discovery systems still match the words of a query against the words in title, subject, keyword and abstract fields and rank the results by term statistics, a family of methods of which BM25 remains the best-known representative (Robertson & Zaragoza, 2009).

Lexical matching serves known-item searching well, but it struggles in three situations that are common in academic work. The first is vocabulary mismatch: a user who asks for ‘low-resource language processing for Telugu’ may miss relevant theses catalogued under broader headings that never use the phrase ‘low-resource’. The second is exploratory or interdisciplinary searching, in which the user cannot yet name the concepts, authors or journals that matter. The third is conversational searching, in which an information need is refined over several turns and each query depends on the previous one. In each case the difficulty lies not in access to content but in navigation: connecting a particular user, with a particular purpose, to the resources that will serve it, and explaining why they are relevant.

These difficulties are sharpened in Indian higher education. Libraries in universities and autonomous colleges typically manage records across an integrated library system such as Koha, an institutional repository such as DSpace, national electronic theses and dissertations services, and several subscribed and open e-resource platforms, each with its own metadata conventions and search interface. The result is a fragmented discovery experience in which relationships between resources — that a thesis in the repository cites an article in a subscribed journal, or that two departments work on overlapping problems — remain invisible to users.

1.2 Knowledge graphs and generative AI as complementary technologies

Two families of technology now offer a way to address these limitations. Knowledge graphs (KGs) represent entities — resources, people, organisations, concepts — as nodes and the typed relationships among them as edges, so that both explicit facts and inferred connections can be queried (Hogan et al., 2021; S. Ji et al., 2022). Libraries have long practised the principles on which KGs depend, through authority control, controlled vocabularies and, more recently, linked data (Hallo et al., 2016). Large language models (LLMs), the engines of generative AI, add a complementary capability: they interpret natural-language questions, sustain multi-turn dialogue and produce fluent summaries. LIS scholarship quickly recognised their potential for reference, discovery and technical services (Adetayo, 2023; C. Cox & Tzoc, 2023; Lund & Wang, 2023). The two technologies are complementary because each is weak where the other is strong. LLMs are prone to hallucination, producing content that is fluent but unsupported or false (Huang et al., 2025; Z. Ji et al., 2023). For libraries, the most damaging form is the fabricated reference: Walters and Wilder (2023) found that 55% of the citations produced by GPT-3.5 in their sample, and 18% of those produced by GPT-4, did not correspond to real works. Retrieval-augmented generation (RAG) mitigates this by conditioning generation on documents retrieved from a trusted collection (Lewis et al., 2020), and KGs can further constrain and explain what is retrieved (Edge et al., 2024; Pan et al., 2024). Conversely, KGs are costly to build and maintain, and LLMs can assist with the entity and relation extraction on which construction depends (Pan et al., 2024). An academic library that combines the two can, in principle, offer conversational search in which every statement is traceable to a record the library holds.

1.3 Research gap

Despite this promise, three gaps limit what academic libraries can currently draw from the literature. First, LIS research on generative AI has been dominated by perception surveys, opinion pieces and conceptual reviews (A. M. Cox et al., 2019; Hervieux & Wheatley, 2021; Panda & Kaur, 2023). Where prototypes exist,

they have mostly answered questions about library services from website content rather than retrieving from bibliographic collections (Lappalainen & Narayanan, 2023); Bevara et al. (2025) outline a RAG architecture for academic library search, but without KG integration or a controlled comparison against existing retrieval. Second, research on KG–LLM integration in computer science (Edge et al., 2024; Oelen et al., 2024; Pan et al., 2024) is seldom designed around library metadata standards, library workflows or the role of librarians in curating the knowledge on which retrieval depends. Third, scholarly recommender systems have usually been evaluated on accuracy alone (Beel et al., 2016), neglecting the diversity, novelty and serendipity that matter to academic discovery (Kaminskas & Bridge, 2016) and the explainability, trust and privacy considerations that shape whether users will adopt AI-driven services.

This conceptual paper addresses these gaps by designing KG-LibRAG, an integrated knowledge graph and generative AI framework for intelligent information retrieval and personalised resource recommendation in academic libraries, and by specifying how it should be evaluated. It does not report empirical results. Instead, it derives the framework from the literature and from design science principles, states testable propositions and sets out an evaluation protocol for subsequent empirical work. Its central research question is: *How can knowledge graphs and generative AI be integrated to develop an intelligent, context-aware information retrieval and personalised resource recommendation framework for academic libraries?*

1.4 Objectives

The paper pursues six objectives:

- O1. To examine the applications of knowledge graphs in semantic information retrieval in academic libraries.
- O2. To investigate the potential of generative AI and large language models for academic information discovery.
- O3. To identify the factors influencing personalised academic-resource recommendation.
- O4. To develop an integrated knowledge graph–generative AI conceptual framework for intelligent library information retrieval.
- O5. To specify testable propositions and an evaluation protocol for assessing the framework in terms of retrieval relevance, personalisation, contextual understanding and recommendation quality.
- O6. To identify the technological, ethical, privacy and professional challenges associated with implementing the framework in academic libraries.

1.5 Research questions

- RQ1. What are the major applications of knowledge graphs in academic information retrieval?
- RQ2. How is generative AI being applied to academic information discovery and recommendation?
- RQ3. What limitations exist in conventional academic-library retrieval systems?
- RQ4. How can knowledge graphs and generative AI be integrated to improve semantic and contextual retrieval?
- RQ5. How should the contribution of such a framework to retrieval relevance and personalised resource recommendation be evaluated?
- RQ6. What ethical, privacy, transparency and professional issues arise from AI-driven personalisation in academic libraries?

1.6 Contributions

The paper makes four principal contributions. First, it proposes a library-domain KG ontology that reuses established bibliographic and knowledge-organisation vocabularies and extends them with the course, research-interest and provenance relations needed for personalised discovery. Second, it specifies a hybrid KG–generative AI architecture in which lexical, dense and graph retrieval are fused, generation is constrained to retrieved evidence, and every cited identifier is verified against the collection before an answer is shown. Third, it provides a personalised recommendation mechanism with path-based explanations. Fourth, it states five research propositions and an evaluation protocol — comparison

conditions, metrics, sample-size targets, an analysis plan and a questionnaire — that allow the framework to be tested against conventional retrieval and the results to be compared across institutions. A further practical contribution is a staged adoption roadmap for libraries that run open-source systems.

1.7 Structure of the paper

Section 2 describes the research approach. Section 3 reviews the literature and synthesises the research gaps. Section 4 sets out the theoretical foundation, design requirements and research propositions. Section 5 presents the KG-LibRAG framework, and Section 6 the protocol for its evaluation. Section 7 discusses the theoretical contribution of the framework, Section 8 its implications for academic libraries, and Section 9 the limitations of the study and an agenda for future research. Section 10 concludes.

2. Research approach

2.1 A conceptual design science approach

The paper follows a design science research (DSR) paradigm, in which knowledge is produced by building and evaluating artefacts that address important organisational problems (Hevner et al., 2004). Hevner et al. (2004) recognise constructs, models and methods, as well as working instantiations, as legitimate design artefacts; this paper contributes artefacts of the first three kinds — an ontology, an architecture and an evaluation method — in advance of an instantiation. It is organised around the six activities of the design science research methodology (DSRM): problem identification and motivation, definition of solution objectives, design and development, demonstration, evaluation and communication (Peppers et al., 2007). Table 1 shows how each activity is treated in this paper and which parts are deferred to subsequent empirical work.

Table 1. Treatment of design science activities in this paper

DSRM activity	Treatment in this paper	Deferred to empirical work	Objectives and RQs
Problem identification and motivation	Integrative review of the literature (Sections 1 and 3)	PRISMA 2020 systematic review to confirm and quantify the gaps	O1–O3; RQ1–RQ3
Definition of solution objectives	Design requirements R1–R8 derived from the review and from sociotechnical theory (Section 4.2)	Validation of requirements with librarians and users	O3, O6; RQ3, RQ6
Design and development	Five-layer architecture, ontology, retrieval, generation and recommendation mechanisms (Section 5)	Implementation of a prototype over library metadata	O4; RQ4
Demonstration	Illustrative scenario (Section 5.9)	Demonstration on a live collection	O4; RQ4
Evaluation	Propositions and evaluation protocol (Sections 4.3 and 6)	Offline and user evaluation	O5, O6; RQ5, RQ6
Communication	This paper	Publication of results, code and test collection	All

Source: Author's own work, based on Peppers et al. (2007)

2.2 Literature base

The literature on which the framework rests was assembled through an integrative review rather than a systematic one. Sources were identified through literature searching and citation chasing from key surveys, and were drawn from six bodies of work: lexical and semantic information retrieval; knowledge graphs and linked data in libraries; generative AI in academic libraries; hallucination, grounding and retrieval-augmented generation; the integration of knowledge graphs and LLMs; and personalised scholarly recommendation. Priority was given to peer-reviewed surveys, foundational methods and reported library

implementations published from 2015 onward, together with earlier foundational works in information retrieval, information systems and technology acceptance. Preprints were used only where they document methods or infrastructures that have been widely adopted. Sources were selected for their conceptual contribution to the design rather than for completeness, so the review does not claim exhaustive coverage; a systematic review is the first item of the research agenda in Section 9.2.

Each design decision in Section 5 is traced to the literature from which it derives, so that the reasoning behind the framework can be examined and challenged. The propositions in Section 4.3 are stated in falsifiable form, and the protocol in Section 6 specifies how each would be tested.

3. Literature review

The review is organised around the six themes on which the framework draws; Section 2.2 describes how the literature was assembled. It closes with a synthesis of research gaps (Section 3.7), from which the design requirements in Section 4.2 are derived.

3.1 From lexical to semantic information retrieval

Probabilistic lexical models, of which BM25 is the most widely deployed, score documents by the frequency and rarity of query terms and remain strong, efficient baselines (Robertson & Zaragoza, 2009). Their limitation is that they operate on surface forms: synonyms, broader and narrower concepts and paraphrases are invisible unless they are added through query expansion or controlled vocabulary.

Dense retrieval addresses this by encoding queries and documents as vectors in a shared semantic space. Dense passage retrieval showed that learned embeddings could outperform BM25 on open-domain question answering (Karpukhin et al., 2020), and Sentence-BERT made semantically meaningful sentence embeddings efficient enough for large-scale similarity search (Reimers & Gurevych, 2019). For scholarly documents, SPECTER used citation relationships as a training signal to produce document-level embeddings suited to recommendation and classification (Cohan et al., 2020), an early indication that relational structure improves semantic representation. Because lexical and dense methods fail in different ways, hybrid systems commonly merge their ranked lists; reciprocal rank fusion (RRF) is a simple, unsupervised method that outperformed more complex alternatives in its original evaluation (Cormack et al., 2009).

Evaluation practice in information retrieval (IR) also shapes the evaluation protocol proposed in Section 6. Graded relevance measures such as normalised discounted cumulative gain (nDCG) reward systems that place highly relevant items near the top of the list (Järvelin & Kekäläinen, 2002), while interactive IR research emphasises that offline metrics must be complemented by studies of how users experience a system (Kelly, 2009). Neither lexical nor dense retrieval, however, captures the explicit relational semantics — authorship, citation, subject hierarchy, institutional affiliation — that librarians curate. This is the contribution of knowledge graphs.

3.2 Knowledge graphs and linked data in libraries

A knowledge graph is a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between them (Hogan et al., 2021). Surveys distinguish four technical concerns: knowledge representation, knowledge acquisition through entity and relation extraction, knowledge-graph completion, and the application of the graph to tasks such as question answering and recommendation (S. Ji et al., 2022).

Libraries are natural producers of such graphs. Authority files, classification schemes and subject headings already encode entities and relations, and the library linked-data movement has sought to publish them in machine-readable form. The Library of Congress's BIBFRAME initiative provides a linked-data successor to MARC (Library of Congress, n.d.); SKOS offers a standard model for thesauri and subject-heading systems (Miles & Bechhofer, 2009); and DCMI Metadata Terms supply widely used descriptive properties (DCMI Usage Board, 2020). A review of linked-data implementations at five major national and European

libraries found improved visibility and reuse of data but identified the multiplicity of vocabularies and the shortage of tools and expertise as persistent barriers (Hallo et al., 2016).

Beyond individual libraries, open scholarly graphs now provide identifiers and relations to which local systems can link. Wikidata is a collaboratively edited, multilingual knowledge base (Vrandečić & Krötzsch, 2014); OpenAlex indexes scholarly works, authors, venues, institutions and concepts under an open licence (Priem et al., 2022); and the Open Research Knowledge Graph represents the contributions of research papers — problems, methods and results — in machine-actionable form (Jaradeh et al., 2019). Few academic libraries, and very few in India, have yet built local KGs that link their own holdings to these resources for the purpose of discovery. Construction cost, rather than conceptual unfamiliarity, is the principal obstacle, and it is precisely here that LLMs can help.

3.3 Generative AI and large language models in academic libraries

Interest in AI in academic libraries predates generative models. Interviews with library leaders and commentators identified search and resource discovery as a primary area of AI impact but raised concerns about the intelligibility of algorithmic decisions, data quality and the marginal position of libraries in AI development (A. M. Cox et al., 2019). A survey of North American academic librarians found limited shared understanding of AI and mixed attitudes towards its adoption (Hervieux & Wheatley, 2021).

The public release of ChatGPT generated a rapid body of LIS commentary. Lund and Wang (2023) outlined the technology and its implications for academia and libraries; Adetayo (2023) reviewed its potential for reference, cataloguing, classification and collection development and concluded that it should complement rather than replace librarians; Panda and Kaur (2023) judged ChatGPT-based chatbots a viable alternative to knowledge-base chatbots while noting risks of bias; C. Cox and Tzoc (2023) discussed implications for academic library services; and Lund et al. (2023) examined the ethics of LLM-written research in scholarly publishing.

Reported implementations remain few. Zayed University Library's Aisha chatbot combined the ChatGPT API with embeddings of library web pages and guides to answer service questions outside staffed hours (Lappalainen & Narayanan, 2023). Bevara et al. (2025) analysed the technical requirements of RAG for academic library search, including embedding pipelines, vector databases and middleware for integration with existing systems. Taken together, this literature establishes demand and plausibility, but it offers little controlled evidence on retrieval effectiveness over bibliographic collections and rarely treats structured knowledge as a source of grounding.

3.4 Hallucination, grounding and retrieval-augmented generation

Hallucination is the generation of content that is nonsensical or unfaithful to the provided source (Z. Ji et al., 2023). More recent taxonomies distinguish factuality hallucinations, which conflict with verifiable facts, from faithfulness hallucinations, which depart from the user's instructions or the supplied context (Huang et al., 2025). Fabricated references are a particularly serious instance for libraries: Walters and Wilder (2023) found that a majority of GPT-3.5 citations and nearly one in five GPT-4 citations in their sample were fabricated, and that many real citations contained substantive errors.

RAG couples a retriever with a generator so that the model conditions its output on documents drawn from an external, updatable source (Lewis et al., 2020). Subsequent work has moved from naive retrieve-then-read pipelines to advanced and modular designs that add query rewriting, re-ranking and iterative retrieval (Gao et al., 2023). Evaluation has correspondingly broadened from answer accuracy to faithfulness to the retrieved context, answer relevance and context relevance, which can be estimated with frameworks such as RAGAs (Es et al., 2024). For an academic library, the implication is that every statement a discovery system generates should be traceable to a record in its own collection, and that traceability should be checked rather than assumed.

3.5 Integrating knowledge graphs and large language models

Pan et al. (2024) propose a roadmap with three modes of integration: KG-enhanced LLMs, in which graphs inform pre-training, inference or interpretation; LLM-augmented KGs, in which language models assist with embedding, completion, construction and question answering over graphs; and synergised systems in which the two reason together. GraphRAG illustrates the synergy: an LLM extracts an entity graph from a corpus, community detection partitions it, and pre-generated community summaries allow the system to answer global sensemaking questions that conventional vector RAG handles poorly (Edge et al., 2024). In the scholarly domain, ORKG ASK combines vector search, LLMs and a scholarly KG in a neuro-symbolic search and exploration system (Oelen et al., 2024).

These systems demonstrate technical feasibility, but they are designed for open corpora or research-contribution graphs rather than for a library's own holdings, licences and user community. None treats the library's controlled vocabularies as a first-class structure, and none positions the librarian as curator of the knowledge on which generation depends.

3.6 Personalised recommendation in scholarly contexts

A comprehensive survey of research-paper recommender systems found that content-based filtering dominated the field and that evaluations were often irreproducible and disconnected from user needs (Beel et al., 2016). KG-based recommenders address two persistent problems — data sparsity and cold start — by using the graph as side information, and they offer a natural route to explanation because a recommendation can be justified by the path that connects user and item (Guo et al., 2022). Models such as the knowledge graph attention network propagate user preferences across high-order graph connections to improve recommendation quality (Wang et al., 2019).

Accuracy is not the only quality that matters. Diversity, novelty, serendipity and coverage — the 'beyond-accuracy' objectives — influence how users perceive recommendations and whether a system narrows or widens their exposure (Kaminskas & Bridge, 2016). These objectives align closely with the library's mission of supporting discovery across the whole collection rather than amplifying already popular items. Personalisation also raises questions of privacy and confidentiality, long-standing commitments of library ethics that responsible AI practice in libraries must uphold (Padilla, 2019).

3.7 Synthesis and research gaps

Table 2 synthesises the gaps identified in the review and the design response of KG-LibRAG to each.

Table 2. Research gaps and the corresponding design responses of KG-LibRAG

Research gap	Supporting literature	Response in KG-LibRAG
Generative AI studies in LIS are largely conceptual or perception-based	A. M. Cox et al. (2019); Hervieux & Wheatley (2021); Adetayo (2023); Panda & Kaur (2023)	An implementable architecture and a controlled four-condition evaluation protocol
LLM prototypes in libraries answer from web content rather than bibliographic collections	Lappalainen & Narayanan (2023); Bevara et al. (2025)	Retrieval over integrated library system and repository metadata with evidence-linked answers
KG-LLM systems are not designed around library metadata or librarian curation	Pan et al. (2024); Edge et al. (2024); Oelen et al. (2024)	Library ontology reusing BIBFRAME, SKOS and DCMI terms; librarian-in-the-loop validation
Hallucinated citations undermine trust in generated answers	Z. Ji et al. (2023); Walters & Wilder (2023)	Generation restricted to retrieved records; post-generation citation verification
Recommenders are evaluated on accuracy alone, with little attention to explanation or privacy	Beel et al. (2016); Kaminskas & Bridge (2016); Guo et al. (2022)	Beyond-accuracy metrics, path-based explanations, consent-based pseudonymous profiles

Research gap	Supporting literature	Response in KG-LibRAG
Technology-centred framing neglects organisational and ethical conditions of adoption	Bostrom & Heinen (1977); Padilla (2019)	Sociotechnical design and evaluation of people, organisation and ethics alongside technology

Source: Author's own work

4. Theoretical foundation, design requirements and propositions

4.1 A sociotechnical perspective

The framework is grounded in sociotechnical systems theory, which holds that organisational information systems succeed only when their technical and social subsystems are designed jointly, and that systems optimised for technical performance alone tend to fail in use (Bostrom & Heinen, 1977). Sociotechnical approaches have since been developed into methods for engineering complex systems in which people, processes and technology interact (Baxter & Sommerville, 2011). An AI discovery system in an academic library is such a system: its outputs depend on metadata created by cataloguers, its acceptability on the trust of students and faculty, its sustainability on institutional policy and staff competence, and its legitimacy on adherence to professional ethics. Table 3 summarises the four dimensions and where the framework addresses each.

Table 3. Sociotechnical dimensions addressed by the framework

Dimension	Relevant issues	Where the framework addresses it
Technology	Knowledge graphs, generative AI and LLMs, embeddings, retrieval and recommendation algorithms	Layers 2–5 (Section 5); retrieval, recommendation and faithfulness metrics (Section 6.3)
People	Students, research scholars, faculty, librarians and research-support professionals; AI literacy; trust	Librarian-in-the-loop curation (Section 5.7); proposed user study (Section 6.5)
Organisation	Library policies, digital infrastructure, staff competencies, governance, implementation readiness	Non-invasive integration (Section 5.2); adoption roadmap (Section 8.3)
Ethics	Privacy, bias, explainability, copyright, transparency, hallucination, responsible AI use	Governance function (Section 5.7); faithfulness and bias audits; evaluation safeguards (Section 6.7)

Source: Author's own work, drawing on Bostrom and Heinen (1977) and Baxter and Sommerville (2011)

At the level of individual users, the framework draws on technology-acceptance research. The technology acceptance model identifies perceived usefulness and perceived ease of use as determinants of intention to use a system (Davis, 1989), and the unified theory of acceptance and use of technology consolidates related constructs into performance expectancy, effort expectancy, social influence and facilitating conditions (Venkatesh et al., 2003). With generative AI, users cannot inspect the reasoning behind an answer, so trust becomes an additional and potentially decisive construct. The proposed user-evaluation model (Figure 1) therefore treats perceived relevance and ease of use as antecedents drawn from these theories and adds personalisation and trust as the constructs of primary interest.

4.2 Design requirements

Eight design requirements follow from the gaps in Table 2 and the dimensions in Table 3. The system should support (R1) interpretation of natural-language and multi-turn queries, because conversational and exploratory searching are where lexical retrieval is weakest; (R2) retrieval that exploits subject, authorship and citation relations, because these are the relations that libraries curate and that lexical and dense methods ignore (Hogan et al., 2021); and (R3) answers in which every statement is linked to a collection record, given the documented prevalence of fabricated citations (Walters & Wilder, 2023). It should offer (R4) personalised recommendations accompanied by human-readable explanations (Guo et al., 2022) and observe (R5) privacy by design, with consent-based and pseudonymous user profiles, in keeping with library

commitments to confidentiality (Padilla, 2019). Organisationally, it should achieve (R6) integration with existing open-source library systems without replacing them, (R7) librarian oversight of the knowledge graph and of system behaviour (A. M. Cox et al., 2019), and (R8) reproducibility through logging of models, prompts and configurations.

4.3 Research propositions

Because the framework has not yet been implemented, its expected effects are stated as propositions rather than as tested hypotheses. Each is formulated so that it can be refuted, and Section 6 specifies the comparison and test through which it would be examined. P1–P3 concern system-level comparisons between retrieval conditions; P4 and P5 concern the user-evaluation model shown in Figure 1.

P1. Knowledge graph-based retrieval yields higher retrieval relevance than conventional keyword-based retrieval, with the largest gains for exploratory and interdisciplinary queries. KG-based expansion should reduce vocabulary mismatch by adding related concepts, authors and subjects to the query, whereas BM25 is already strong for known-item searching (Hogan et al., 2021; Robertson & Zaragoza, 2009).

P2. Generative AI-assisted retrieval increases users' perceived search effectiveness relative to keyword-based retrieval. Natural-language interaction, query rewriting and synthesised answers should reduce the effort of formulating and refining queries (Lewis et al., 2020; Venkatesh et al., 2003). Because perceived effectiveness is not evidence of accuracy, the faithfulness of generated answers is assessed separately (Section 6.3).

P3. The integration of knowledge graphs and generative AI improves personalised resource recommendation, in both accuracy and beyond-accuracy terms, relative to either component alone. Graph paths link user interests to resources that lexical similarity misses, while the LLM interprets context and explains recommendations (Guo et al., 2022; Kaminskas & Bridge, 2016; Pan et al., 2024).

P4. Perceived personalisation is positively associated with users' satisfaction with AI-based academic-library information retrieval. Recommendations that reflect a user's course and research interests should reduce search effort and increase perceived value.

P5. Perceived trust is positively associated with users' intention to adopt AI-based library discovery systems. Where users cannot inspect how an answer was produced, confidence that cited sources are genuine should condition continued use (Davis, 1989; Venkatesh et al., 2003).

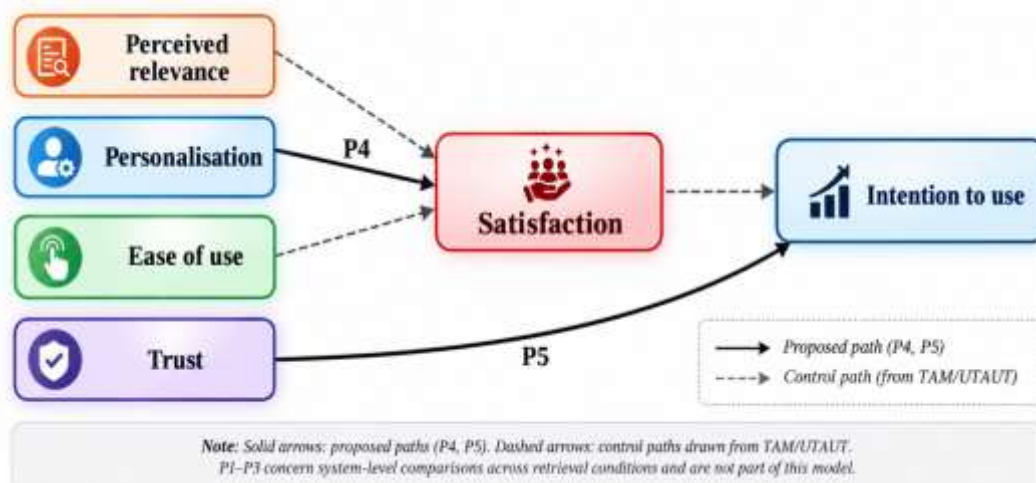


Figure 1. Proposed research model for the user evaluation

Source: Author's own work, drawing on Davis (1989) and Venkatesh et al. (2003)

5. The proposed KG-LibRAG framework

5.1 Overview

KG-LibRAG is organised in five layers, with a cross-cutting governance function and a feedback loop (Figure 2). Data flow upward: library data sources (Layer 1) are transformed into a library knowledge graph (Layer 2); a generative AI layer (Layer 3) interprets user queries against that graph; an intelligent retrieval

layer (Layer 4) combines lexical, dense and graph retrieval; and a personalised recommendation layer (Layer 5) tailors and explains results for each user. User feedback returns to the profile and, after librarian review, to the graph. The end-to-end research model can be summarised as: academic library data → knowledge graph → semantic and vector retrieval → generative AI → user context and research profile → personalised retrieval → explainable resource recommendations → user feedback → continuous system improvement. The feedback loop makes the framework an adaptive discovery system rather than a static AI search interface. The layers and the governance function map directly onto the design requirements in Section 4.2: Layer 3 addresses R1; Layer 4 addresses R2 and R3; Layer 5 addresses R4; the non-invasive design of Layer 1 addresses R6; and the governance function addresses R5, R7 and R8.

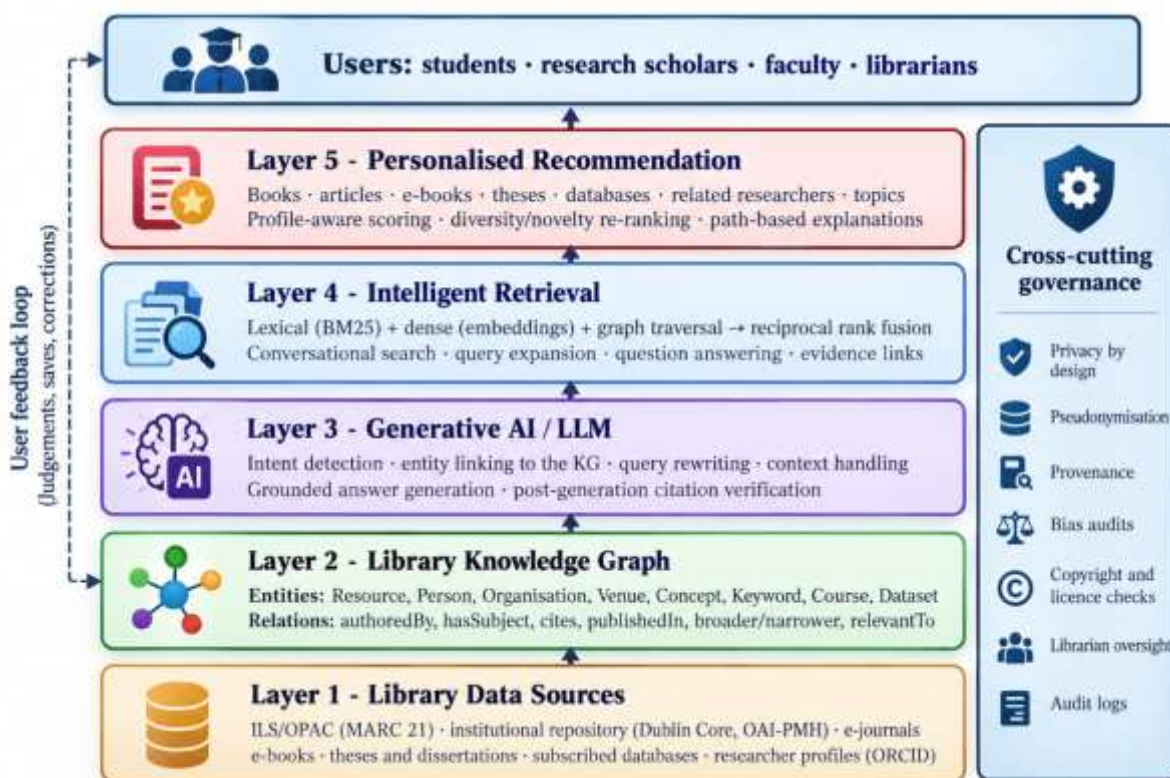


Figure 2. The five-layer KG-LibRAG architecture

Source: Author's own work

5.2 Layer 1: library data sources and ingestion

Layer 1 ingests metadata from systems that an academic library already runs: the integrated library system and OPAC (MARC 21 records), the institutional repository (Dublin Core through OAI-PMH), theses and dissertations, e-book and e-journal title lists supplied by vendors, subscribed databases and researcher profiles. A crosswalk maps each source into a common internal schema. The layer is deliberately non-invasive: KG-LibRAG reads from existing systems and writes nothing back to them, so libraries can adopt it without migrating their catalogue. This matters in institutions that rely on open-source systems such as Koha and DSpace and have limited technical staff. Records are refreshed incrementally on a schedule that matches cataloguing activity.

Table 4 lists the metadata fields on which the framework draws and their role in the knowledge graph. Where records carry DOIs, they can be enriched with open citation and concept data from OpenAlex (Priem et al., 2022).

Table 4. Metadata fields and their role in the knowledge graph

Field	Typical source	Role in KG-LibRAG
Title; abstract	MARC 245, 520; dc.title, dc.description.abstract	Lexical and dense indexing; entity and relation extraction
Author(s)	MARC 100, 700; dc.contributor.author; ORCID where available	Person nodes; authoredBy relations; name disambiguation
Subject; keywords	MARC 650, 653; dc.subject	Concept and Keyword nodes linked to controlled vocabulary
Classification	MARC 082 (DDC) or local scheme	Broader and narrower subject structure
Publication year; publisher	MARC 264/260; dc.date.issued, dc.publisher	Filtering, recency features, Organisation nodes
Journal or venue	MARC 773; e-resource title lists	Venue nodes; publishedIn relations
DOI; ISBN	MARC 020, 024; dc.identifier	Deduplication; linking to open scholarly graphs
Citation information	OpenAlex (via DOI); repository reference lists	cites relations; citation-aware ranking and recommendation
Department; programme (theses)	Repository collection structure; local fields	Organisation and Course nodes; relevantTo relations

Note: MARC tags and Dublin Core elements are indicative; local cataloguing practice varies. Source: Author's own work

5.3 Layer 2: the library knowledge graph

Layer 2 is a property graph whose ontology reuses established vocabularies wherever possible — BIBFRAME for works and instances (Library of Congress, n.d.), SKOS for subject concepts (Miles & Bechhofer, 2009) and DCMI terms for descriptive properties (DCMI Usage Board, 2020) — and extends them only where discovery requires it. Table 5 lists the core classes and relations.

Table 5. Core classes and relations of the library ontology

Class	Examples	Key relations
Resource	Book, article, e-book, thesis, dataset, database	authoredBy → Person; hasSubject → Concept; publishedIn → Venue; cites → Resource; relevantTo → Course
Person	Author, supervisor, researcher	affiliatedWith → Organisation; researchesOn → Concept; coAuthorWith → Person
Organisation	University, college, department, publisher, funder	partOf → Organisation
Venue	Journal, conference, book series	coversSubject → Concept
Concept	Subject heading, classification class, research area, method, task	broader, narrower, related → Concept (SKOS)
Keyword	Uncontrolled author or cataloguer terms	mapsTo → Concept
Course	Programme or course of study; syllabus unit	offeredBy → Organisation; covers → Concept
UserProfile	Consenting user's pseudonymous interests and interaction summary	interestedIn → Concept; interactedWith → Resource (aggregated)

Note: UserProfile nodes are held in a separate, access-controlled partition and are never exposed in generated answers. Source: Author's own work

Two extensions are distinctive. The Course class links resources to programmes and syllabus units, so that recommendations can follow the curricula that shape most academic information needs, particularly at undergraduate level. The UserProfile class stores only consented, pseudonymous summaries of interests and interactions, kept apart from bibliographic data. Every edge carries provenance and a confidence value, so

that catalogued relations can be distinguished from machine-extracted ones and weighted accordingly during retrieval. LLMs assist with graph construction by proposing relations from abstracts, following the LLM-augmented KG pattern (Pan et al., 2024), but proposed relations below a confidence threshold enter a librarian review queue before they are used.

Records are de-duplicated using identifiers and fuzzy matching of title and author strings, normalised for names, dates and diacritics, and tagged by language. Subject strings are reconciled to the library's controlled vocabulary and represented as SKOS concepts. Entity and relation extraction combines rule-based mapping of structured fields with LLM-assisted extraction of methods, tasks and research areas from abstracts; a stratified sample of extracted triples is validated by librarians, and the precision measured on that sample sets the confidence threshold below which machine-extracted relations are held for librarian review. Author names are disambiguated using ORCID identifiers where present and co-authorship and affiliation evidence otherwise. Every triple carries provenance indicating its source record and whether it was catalogued, harvested or machine-extracted.

5.4 Layer 3: the generative AI layer

Layer 3 performs four functions. First, it detects the intent of a query: known-item lookup, topical search, exploratory question or conversational follow-up. Second, it links mentions in the query to KG entities — concepts, authors, venues and courses — so that retrieval can operate on identifiers rather than strings. Third, it rewrites and expands the query using the KG neighbourhood of the linked entities (narrower concepts, preferred and alternative labels, closely related methods) and the context of earlier turns. Fourth, after retrieval, it generates an answer under a grounding prompt that restricts it to the supplied records, requires each claim to cite a record identifier and instructs it to say when the collection does not contain an answer (Appendix B). The layer is model-agnostic; for reasons of privacy and cost, an open-weight model hosted on institutional infrastructure is preferred where its quality is adequate.

5.5 Layer 4: intelligent retrieval

Layer 4 implements the hybrid retrieval pipeline shown in Figure 3. Three retrievers run in parallel over the same collection. A lexical retriever applies BM25 to the expanded query (Robertson & Zaragoza, 2009). A dense retriever compares query and record embeddings produced by sentence- and document-level encoders (Cohan et al., 2020; Reimers & Gurevych, 2019). A graph retriever starts from the linked entities and traverses the KG to a bounded depth, collecting resources connected through subject, authorship, citation and course relations and scoring them by path type, path length and edge confidence. For broad sensemaking questions, it can also draw on pre-computed summaries of densely connected communities in the graph (Edge et al., 2024).

The three ranked lists are merged by reciprocal rank fusion, in which the fused score of a record d is the sum over retrievers r of $1 / (k + \text{rank}_r(d))$, with $k = 60$ (Cormack et al., 2009). RRF requires no training data, which suits libraries without large relevance-labelled logs. The top-ranked records, together with the KG paths that connect them to the query entities, are passed to Layer 3 for grounded generation. A verification step then checks every identifier cited in the draft answer against the retrieved set and the catalogue; any claim whose citation is missing or does not resolve is removed or flagged before the answer is displayed. This step targets directly the fabricated-citation problem documented by Walters and Wilder (2023). The answer is presented alongside the ranked result list, so that users can move between a synthesised overview and the underlying records.

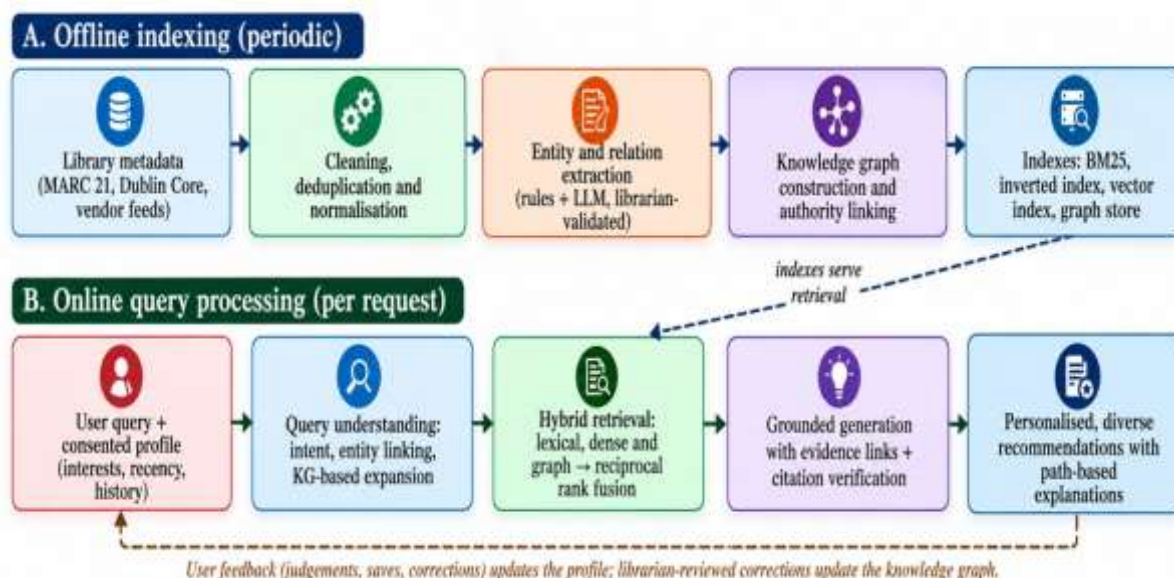


Figure 3. Offline indexing and online query-processing pipeline of KG-LibRAG

Source: Author's own work

5.6 Layer 5: personalised recommendation

Layer 5 generates recommendations for books, articles, e-books, theses, databases, related researchers and research topics. Candidate items are drawn from the KG by following paths from the user's profile — for example, from a user's interest to a concept, from that concept to a narrower concept, and from the narrower concept to the resources that carry it as a subject — and from the dense neighbourhood of items with which the user has recently engaged. Each candidate d for user u , in the context of query q , receives a score of the form:

$$score(u, d | q) = \lambda_1 \cdot rel(q, d) + \lambda_2 \cdot aff(u, d) + \lambda_3 \cdot prox_KG(u, d) + \lambda_4 \cdot nov(d)$$

where rel is the fused retrieval relevance, aff the similarity between the item and the user's interest profile, $prox_KG$ a path-based proximity in the graph and nov an inverse-popularity term; the weights λ are tuned on a development split of held-out interactions, with equal weights as a transparent starting point where tuning data are lacking. The ranked list is then re-ranked to balance relevance against diversity, so that the final list does not collapse onto a single sub-topic (Kaminskas & Bridge, 2016). For cold-start users with no interaction history, declared interests and course membership provide the entry points into the graph (Guo et al., 2022).

Every recommendation carries an explanation generated from its highest-scoring path — for example, 'recommended because it addresses sentiment analysis, a narrower topic of your stated interest in Telugu natural language processing, and is cited by two theses from your department'. Path-based explanations make the reasoning inspectable, give users grounds for trust or scepticism, and allow librarians to diagnose poor recommendations.

5.7 Governance and the feedback loop

A governance function spans all layers: consent management and pseudonymisation of profiles; provenance and confidence on every KG edge; licence checks before any content is indexed or sent to a model; periodic audits of recommendation bias across disciplines, languages, publishers and publication years; audit logs of prompts, model versions and retrieved evidence; and a librarian dashboard for reviewing machine-extracted relations and user-reported errors. User feedback — explicit relevance judgements, saves and corrections — updates the user profile directly, whereas corrections to bibliographic facts are routed to librarians before they alter the graph. In this way the system improves with use, while authority over the library's knowledge remains with the profession.

5.8 Reference implementation

The framework is technology-agnostic, but Table 6 lists candidate components for a reference implementation. The architecture is deliberately modular, so that each component can be replaced as tools evolve without changing the framework.

Table 6. Candidate technology stack for a reference implementation

Component	Candidate technologies	Role
Programming and orchestration	Python; a RAG orchestration library or custom pipeline	Data processing, pipeline control, experiment scripts
Knowledge graph store	Neo4j (property graph, Cypher) or an RDF triple store with SPARQL	Storage and traversal of the library knowledge graph
Lexical index	Elasticsearch or OpenSearch (BM25)	Baseline and hybrid lexical retrieval
Vector index	A vector database or library (e.g. FAISS, pgvector, Qdrant)	Dense retrieval over embeddings
Embedding models	Sentence-transformer and scholarly-document encoders	Semantic representation of queries, records and KG entities
Large language model	An open-weight instruction-tuned model on institutional infrastructure (for example, from the Llama or Mistral families) or a commercial API; the exact model and version are recorded in any evaluation	Query understanding, grounded generation, explanation
Evaluation	trec_eval-compatible tooling; RAGAs; SPSS; SmartPLS	Offline metrics, faithfulness, statistical analysis
Interface	Web front end linked to the OPAC	Pilot deployment and user study

Note: Hosting an open-weight model locally keeps queries and profiles on institutional infrastructure; if a commercial API is used, its data-processing terms must be reviewed against the safeguards in Section 6.7. Source: Author's own work

5.9 Illustrative scenario

The following scenario is hypothetical: it illustrates how the layers are intended to interact and does not report the output of an implemented system. A research scholar in computer science asks: ‘What recent work is there on transformer models for Telugu text classification, and what should I read first?’ Layer 3 identifies an exploratory topical need and links ‘Telugu’ to a language concept, ‘text classification’ to a task concept and ‘transformer models’ to a method concept. KG expansion adds narrower tasks such as sentiment analysis and named-entity recognition, together with alternative labels used in local cataloguing. In Layer 4, BM25 retrieves records that contain the literal terms, the dense retriever finds semantically similar abstracts that use different vocabulary, and the graph retriever surfaces theses from the scholar’s department that are linked to the same concepts, as well as the articles they cite. After fusion, the LLM drafts a short overview in which every statement cites a record identifier, and the verification step confirms that each identifier resolves. Layer 5 then orders a reading list, placing a survey article first because of its centrality in the local citation subgraph, and explains each item. If the scholar marks a recommendation as not relevant, the profile is updated; if they report that a thesis is misclassified, the report is queued for a librarian.

6. Proposed evaluation protocol

The framework is not evaluated in this paper. This section specifies a protocol for doing so, so that the propositions in Section 4.3 can be tested and results obtained in different libraries can be compared. The protocol is designed for a single participating institution in the first instance — for example, an Indian university or autonomous college whose library runs Koha and DSpace — and combines an offline evaluation, which measures retrieval, recommendation and generation quality against expert relevance judgements, with a user study, which measures how the framework is experienced (Kelly, 2009).

6.1 Comparison conditions

Four retrieval conditions are compared (Table 7). To isolate the contribution of each component, all conditions share the same collection, pre-processing and user interface, and conditions C3 and C4 use the same LLM and prompt templates.

Table 7. Comparison conditions

Condition	Description	Components
C1 Keyword baseline	Conventional lexical retrieval representing current OPAC and discovery behaviour	BM25 over title, subject, keyword and abstract fields
C2 Knowledge graph-based	Query entities linked to the KG; results from graph traversal and KG-based query expansion	Entity linking; bounded k-hop traversal; BM25 on the expanded query
C3 Generative AI-assisted	LLM rewrites the query; dense retrieval; LLM generates an answer from retrieved records	Query rewriting; dense retrieval; RAG generation
C4 Hybrid (KG-LibRAG)	The full framework	Lexical, dense and graph retrieval fused by RRF; KG-constrained generation; citation verification; personalisation

Note: An ablation of C4 without the citation-verification step is also run, to estimate the contribution of that step to the faithfulness of generated answers. Source: Author's own work

6.2 Collection and test collection

A working collection of 10,000–50,000 records is proposed, drawn from the participating library's integrated library system (MARC 21), its institutional repository (Dublin Core harvested through OAI-PMH) and the title lists of its subscribed e-resources, subject to licensing. Records with DOIs are enriched with open citation and concept data from OpenAlex (Priem et al., 2022). The composition of the collection by resource type and discipline, and the number of records with persistent identifiers and controlled subject headings, are reported, because they bound what graph retrieval can achieve.

A test collection of 50–100 information needs is constructed from anonymised OPAC search logs, reference enquiries and topics supplied by faculty and research scholars. It is stratified across disciplines and across four query types: known-item, topical, exploratory and multi-turn conversational. For each need, the top 20 results from every condition are pooled and judged on a four-point graded scale (0 = not relevant, 1 = marginally relevant, 2 = relevant, 3 = highly relevant) by two subject librarians or domain experts who are blind to the condition that produced each result. Agreement is measured with weighted kappa; a minimum agreement level (for example, $\kappa_w \geq 0.60$) is set in advance, assessors recalibrate on a practice set if it is not reached, and remaining disagreements are adjudicated by a third assessor. The resulting judgements are the ground truth for the retrieval and recommendation metrics.

6.3 Evaluation metrics

Table 8 defines the metrics. Retrieval metrics follow standard IR practice; recommendation metrics include beyond-accuracy objectives; and generation metrics assess whether answers are faithful to the evidence retrieved.

Table 8. Evaluation metrics

Category	Metric	Definition and purpose
Retrieval	Precision	Proportion of retrieved items that are relevant ($\text{grade} \geq 2$)
	Recall	Proportion of relevant items that are retrieved
	F1-score	Harmonic mean of precision and recall, $2PR / (P + R)$
	MRR	Mean over queries of $1 / \text{rank of the first relevant item}$; rewards early relevance

Category	Metric	Definition and purpose
	nDCG@10	DCG@10 / IDCG@10, where $DCG@k = \sum (2^{\text{rel}_i} - 1) / \log_2(i + 1)$; graded ranking quality (Järvelin & Kekäläinen, 2002)
	Precision@5; Recall@10	Precision among the top five results; recall among the top ten results
Recommendation	Precision@K; Recall@K	Accuracy and coverage of the top-K recommendations ($K = 5, 10$)
	Hit rate@K	Proportion of users for whom at least one relevant item appears in the top K
	NDCG@K	Ranking quality of the recommendation list
	Diversity	Intra-list diversity: mean pairwise dissimilarity ($1 - \text{cosine similarity}$) among recommended items (Kaminskas & Bridge, 2016)
	Novelty	Mean self-information of recommended items, $-\log_2 p(i)$, where $p(i)$ is the item's share of loans or views
	Serendipity	Proportion of recommendations judged relevant that fall outside the user's declared interest concepts
Generation	Faithfulness	Proportion of claims in an answer supported by the retrieved records (Es et al., 2024)
	Answer relevance	Degree to which the answer addresses the information need (Es et al., 2024)
	Citation validity	Proportion of cited identifiers that exist in the collection and support the claim to which they are attached
	Hallucination rate	Proportion of answers containing at least one unsupported claim, from a manual audit of a stratified random sample of answers from each generating condition

Source: Author's own work

6.4 Recommendation evaluation

Recommendation is evaluated offline and in the user study. Offline, profiles are built for consenting participants — ideally the same people who take part in the user study — from declared research interests, programme or course, and pseudonymised interaction history (searches, loans and downloads). A time-based split holds out each user's most recent interactions as test items; conditions C2, C3 and C4 generate top-10 recommendations, and C1 is represented by a popularity-plus-keyword baseline. Participants also judge the relevance of their own recommendations, which supplies the ground truth for serendipity.

6.5 User study

A within-subjects study compares the keyword baseline (C1) with the full framework (C4), the two conditions that matter most for an adoption decision; all four conditions are compared in the offline evaluation. Participants are undergraduate and postgraduate students, research scholars and faculty who regularly use the library. Each participant completes eight realistic search tasks drawn from the test collection, four in each condition, with the order of tasks and conditions counterbalanced by a Latin-square design to control learning and fatigue effects. After each condition, participants complete a questionnaire of five-point Likert items (1 = strongly disagree, 5 = strongly agree) measuring perceived relevance, search effectiveness, personalisation, ease of use, trust, satisfaction and intention to use (Appendix A). Items are adapted from established acceptance instruments (Davis, 1989; Venkatesh et al., 2003) and are pilot-tested with a small group of users (for example, 10–15) before the main study. Short semi-structured interviews with a subsample explore reasons for trust or distrust.

The sample size is set by a priori power analysis. For the paired comparison of C1 and C4, detecting a medium effect ($d = 0.5$) with a two-tailed test at $\alpha = 0.05$ and power of 0.80 requires 34 participants. The structural model is estimated on ratings of C4; following the guidance that sample size should reflect the part of the model with the most predictors (Hair et al., 2019), satisfaction, with three predictors, requires 77 participants to detect a medium effect ($f^2 = 0.15$) at the same α and power. A target of 100 participants therefore allows for incomplete responses and exclusions.

6.6 Analysis plan

The choice of tests follows the design and the measurement properties of each variable (Table 9). Because conditions are compared on the same queries and the same participants, paired and repeated-measures tests are used. An omnibus comparison across C1–C4 (repeated-measures ANOVA, or the Friedman test where assumptions are not met) precedes pairwise tests. Normality of per-query differences is checked with the Shapiro–Wilk test, and non-parametric alternatives are used where it is violated. Pairwise comparisons are corrected for multiplicity with the Holm–Bonferroni procedure, and effect sizes are reported alongside p-values. Retrieval results are also reported separately by query type, because P1 predicts that gains differ between known-item and exploratory queries.

The measurement model for the questionnaire is assessed for indicator reliability, internal consistency (Cronbach’s alpha and composite reliability), convergent validity (average variance extracted) and discriminant validity (heterotrait–monotrait ratio) before structural paths are estimated with partial least squares structural equation modelling (PLS-SEM), following current reporting guidance (Hair et al., 2019). Analyses can be carried out in SPSS and SmartPLS at a significance level of $\alpha = 0.05$.

Table 9. Propositions and analysis plan

Proposition	Dependent measure	Comparison or path	Primary test	Alternative
P1	nDCG@10, P@5, MRR (per query)	C2 vs C1	Paired t-test	Wilcoxon signed-rank test
P2	Perceived search effectiveness (composite)	C4 vs C1 (user study); C3 vs C1 where tested	Paired t-test	Wilcoxon signed-rank test
P3	P@10, NDCG@10, diversity, serendipity (per user)	C4 vs C2 and C3	Repeated-measures ANOVA with post hoc tests	Friedman test with Wilcoxon post hoc tests
P4	Satisfaction	Personalisation → Satisfaction	PLS-SEM path coefficient	Multiple regression
P5	Intention to use	Trust → Intention to use	PLS-SEM path coefficient	Multiple regression

Source: Author’s own work

6.7 Ethical safeguards for the evaluation

Before any data are collected, approval for the use of library metadata and for the user study is obtained from the participating institution’s ethics committee. Participation is voluntary and based on informed consent, and participants may withdraw at any time. Search logs and interaction histories are pseudonymised before analysis; no personally identifiable information is used to train or prompt models; and profile data are limited to what personalisation requires and deleted at the end of the study. The source and provenance of every metadata record are documented. Only metadata and abstracts that the library is licensed to process are indexed, and the full texts of licensed resources are not sent to external services.

Generated answers are audited for hallucination and unsupported statements, recommendations are linked to evidence, and recommendation lists are examined for bias in the visibility of resources by discipline, language, publisher and publication age. The model name and version, prompts, retrieval configuration and evaluation protocol are recorded to support reproducibility, and, where licensing permits, the code, query set and relevance judgements are released.

7. Discussion

Because the framework has not yet been evaluated, this section discusses its theoretical position and the evidence that would support or refute each proposition, rather than reporting findings.

7.1 Relational retrieval and the value of curated metadata

The framework rests on the premise that much of the failure of library discovery is relational rather than lexical: the resources a user needs are connected to the query through subject hierarchies, authorship and citation, not only through shared words. If P1 holds, the gain from graph retrieval should be concentrated in exploratory and interdisciplinary queries, where vocabulary mismatch is greatest, and should be small for known-item searches, where BM25 is already strong (Robertson & Zaragoza, 2009). Such a pattern would indicate that the value of cataloguing work — authority control, subject analysis and classification — increases rather than diminishes when AI is introduced, because it is this curated structure that the graph retriever exploits. Conversely, a finding that graph retrieval adds little even for exploratory queries would suggest that local metadata are too sparse or inconsistent to support it, which would itself be an important result for cataloguing policy and for the metadata-readiness stage of the adoption roadmap (Section 8.3).

7.2 Generative AI: fluency and fidelity

Generative AI-assisted retrieval is expected to improve perceived search effectiveness (P2) by reducing the effort of formulating queries and by synthesising results. Yet fluency can mask error. The framework therefore separates two questions: whether users feel better served, and whether what they are told is true. The protocol answers the second question with faithfulness, citation-validity and hallucination measures, and with an ablation that removes the citation-verification step (Section 6.1). A persistent gap between perceived effectiveness and measured faithfulness would signal that users may over-trust fluent answers, echoing the fabricated-citation findings of Walters and Wilder (2023) and the wider hallucination literature (Huang et al., 2025; Z. Ji et al., 2023). Treating verification as an architectural component rather than a prompt instruction is the framework's main response to this risk: a model can be asked to cite only supplied records, but only a check against the catalogue can confirm that it has done so.

7.3 Personalisation, explainability and trust

P3–P5 concern whether personalisation adds value and whether trust governs adoption. KG paths serve both purposes: they generate candidates that similarity-based methods miss, and they supply explanations that users can inspect (Guo et al., 2022). If personalisation is associated with satisfaction (P4) and trust with intention to use (P5), the implication for design is that explanation is not an optional interface feature but a condition of adoption. Beyond-accuracy measures matter here as well: a recommender that is accurate but narrow may satisfy users in the short term while reducing their exposure to the wider collection, which runs against the library's discovery mission (Kaminskas & Bridge, 2016). The framework's diversity-aware re-ranking and novelty term are intended to guard against this, and the protocol measures whether they do.

7.4 The librarian's role and AI literacy

The framework redistributes rather than removes professional work. Librarians curate the vocabularies and authority data on which the graph depends, validate machine-extracted relations, review user-reported errors, audit recommendation bias and help users interpret AI-generated answers. This is consistent with earlier expectations that AI would create roles in data curation, infrastructure building and user guidance in academic libraries (A. M. Cox et al., 2019), and with the view that generative AI should complement rather than replace librarians (Adetayo, 2023). It also implies a substantial competency gap, given the limited shared understanding of AI reported among librarians (Hervieux & Wheatley, 2021). AI literacy, for staff and users alike, becomes part of the infrastructure that makes such a system trustworthy.

7.5 Privacy, bias and copyright

Personalisation depends on data that libraries have traditionally protected. KG-LibRAG therefore limits profiles to consented, pseudonymous summaries, keeps them apart from bibliographic data and never exposes them in generated text. Bias can enter through the collection (under-representation of regional-

language or open-access publications), through the graph (densely catalogued areas attract more paths) and through the model (popularity and language biases acquired in pre-training). The bias audit proposed in Section 6.7 examines visibility by discipline, language, publisher and publication age; any systematic disadvantage should be treated as a collection-development and cataloguing issue as well as an algorithmic one. Copyright and licensing constrain what may be indexed and what may be sent to external models, which is a further argument for hosting models locally and grounding them in metadata and abstracts rather than full text (Padilla, 2019).

8. Implications for academic libraries

8.1 Changing professional roles

Implementing KG-LibRAG calls for competencies that are only partly present in library staffing today: knowledge engineering and ontology maintenance; metadata-to-graph curation and reconciliation with external identifiers; evaluation of AI systems, including the design of test collections and relevance judgements; the design of AI-enabled research-support services; and the governance of data protection and responsible AI. Many of these extend long-standing cataloguing and reference expertise rather than replacing it.

8.2 Staff development and AI literacy

Because these competencies are new, their development must be planned. Continuing professional development programmes — faculty development and refresher courses, consortium-led workshops and online courses — can build them progressively, beginning with AI literacy and the evaluation of generated content and advancing to graph modelling and system evaluation for the staff who will maintain the system. Libraries can also become providers of AI literacy for their users, teaching students and researchers to verify AI-generated answers against their sources.

8.3 A phased adoption roadmap

Table 10 sets out a staged path that allows libraries to adopt the framework incrementally, with each stage delivering value on its own.

Table 10. Phased adoption roadmap for academic libraries

Stage	Focus	Key activities	Readiness indicator
1. Metadata readiness	Data quality	Deduplicate records; enrich with identifiers (DOI, ISBN, ORCID); reconcile subject headings to a controlled vocabulary	Share of records with persistent identifiers and controlled subjects
2. Knowledge graph pilot	Structure	Build the KG for one subject area; librarian validation of extracted relations; graph-based browsing in the OPAC	Extraction precision; staff able to maintain the ontology
3. Grounded conversational search	Retrieval and generation	Hybrid retrieval with evidence-linked answers and citation verification; faithfulness audits	Citation validity and faithfulness above agreed thresholds
4. Consent-based personalisation	Recommendation	Opt-in profiles; path-based explanations; bias audits	User trust and satisfaction; bias-audit results
5. Federation	Scale	Share ontology and graph segments with consortium partners; link to national and open scholarly graphs	Interoperable identifiers and agreed governance

Source: Author's own work

8.4 Policy and governance

Adoption also requires institutional policy: a statement on acceptable AI use in library services; consent and retention rules for personalisation data; criteria for choosing between locally hosted and external models;

procedures for handling reported errors; and periodic reporting of audit results. Such policies turn the ethical commitments set out in Sections 5.7 and 6.7 from project safeguards into standing library practice.

9. Limitations and future research

9.1 Limitations

Several limitations qualify the contribution of this paper. Most fundamentally, the framework has not been implemented or empirically validated, so its propositions express expectations derived from the literature rather than demonstrated effects. The review on which it rests is integrative and purposive rather than systematic, and relevant studies may have been missed. The architecture assumes metadata of reasonable quality and computational infrastructure sufficient for hybrid retrieval and, preferably, local model hosting; both vary widely across academic libraries, and the framework may be difficult to realise where they are weak. Its generative components depend on LLM and embedding technologies that are changing rapidly, so specific components may be superseded, although the modular design is intended to allow them to be replaced without altering the framework. Grounding and verification reduce hallucination but cannot eliminate errors of interpretation within correctly cited records. A knowledge graph built from existing metadata inherits its gaps and biases, and machine-extracted relations retain residual errors even after sample-based validation. Finally, the proposed evaluation begins at a single institution with predominantly English-language metadata, which will limit what can be concluded about other collections and about multilingual discovery.

9.2 An agenda for future research

The first priority is to test the framework. A systematic review following the PRISMA 2020 guideline (Page et al., 2021) would confirm and quantify the gaps identified in Section 3, using a core search string such as the following, adapted to each database and supplemented by broader strings that omit the library or knowledge-graph terms:

```
"knowledge graph*" AND ("generative AI" OR "large language model*" OR LLM)
AND ("information retrieval" OR "recommendation system*")
AND (librar* OR "academic librar*")
```

A prototype built over one institution's Koha and DSpace metadata would then allow the offline evaluation and user study specified in Section 6, and replication in libraries of different sizes, disciplines and metadata quality would show how far the results generalise.

Beyond evaluation, several extensions follow. Multimodal retrieval could extend the framework to images, figures, audio-visual material and manuscripts. Agentic AI, in which the system plans multi-step searches, consults several sources and asks clarifying questions, could support complex research tasks such as literature reviews. Recommendation could be extended to research data, software and funding opportunities, and made citation-aware so that it reflects the structure of scholarly debate rather than topical similarity alone. Research on explainable AI is needed to test which forms of explanation best calibrate user trust. Federated library knowledge graphs, shared across consortia and linked to national services, could spread the cost of construction and enable discovery across institutions. Finally, multilingual retrieval and Indian-language academic discovery — beginning with languages such as Telugu and Hindi, whose scholarly output is poorly served by current discovery systems — represent an important opportunity for Indian academic libraries to lead.

10. Conclusion

Academic library discovery has long depended on matching words, while the knowledge that libraries curate — about subjects, authors, citations and curricula — has remained largely unused at the moment of search. This paper has proposed KG-LibRAG, a conceptual framework that places that knowledge at the centre of generative AI-based discovery. It contributes a library-domain ontology; a hybrid architecture in which lexical, dense and graph retrieval are fused and generation is constrained to verifiable evidence; a

personalised recommendation mechanism with path-based explanations; and five propositions with an evaluation protocol through which the framework can be tested against conventional retrieval. Its contribution is conceptual: the framework is offered as a basis for empirical work, and the protocol is designed to make that work comparable across institutions. The broader argument is that trustworthy AI in academic libraries is a sociotechnical achievement, which depends as much on metadata quality, librarian oversight, user consent and professional competence as on the capability of the models themselves.

Declarations

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests: No Competing Interests.

Ethics approval: Not applicable. This conceptual study did not involve human participants, personal data or library user records. The evaluation proposed in Section 6 will require institutional ethics approval before any data are collected.

Data availability: Data sharing is not applicable to this article, as no new data were created or analysed in this study.

Declaration of generative AI use: In preparing this manuscript, the author used Claude (Anthropic), a large language model, to assist with structuring the manuscript, drafting and editing text, and producing the figures from the author's specifications. The author reviewed and revised all content, checked every reference against its DOI or stable URL, and takes full responsibility for the content of the publication.

References

1. Adetayo, A. J. (2023). Artificial intelligence chatbots in academic libraries: The rise of ChatGPT. *Library Hi Tech News*, 40(3), 18–21. <https://doi.org/10.1108/LHTN-01-2023-0007>
2. Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, 23(1), 4–17. <https://doi.org/10.1016/j.intcom.2010.07.003>
3. Beel, J., Gipp, B., Langer, S., & Breiting, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, 17(4), 305–338. <https://doi.org/10.1007/s00799-015-0156-0>
4. Bevara, R. V. K., Lund, B. D., Mannuru, N. R., Karedla, S. P., Mohammed, Y., Kolapudi, S. T., & Mannuru, A. (2025). Prospects of retrieval augmented generation (RAG) for academic library search and retrieval. *Information Technology and Libraries*, 44(2). <https://doi.org/10.5860/ital.v44i2.17361>
5. Bostrom, R. P., & Heinen, J. S. (1977). MIS problems and failures: A socio-technical perspective. Part I: The causes. *MIS Quarterly*, 1(3), 17–32. <https://doi.org/10.2307/248710>
6. Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2270–2282). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.207>
7. Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 758–759). Association for Computing Machinery. <https://doi.org/10.1145/1571941.1572114>
8. Cox, A. M., Pinfield, S., & Rutter, S. (2019). The intelligent library: Thought leaders' views on the likely impact of artificial intelligence on academic libraries. *Library Hi Tech*, 37(3), 418–435. <https://doi.org/10.1108/LHT-08-2018-0105>
9. Cox, C., & Tzoc, E. (2023). ChatGPT: Implications for academic libraries. *College & Research Libraries News*, 84(3), 99–102. <https://doi.org/10.5860/crln.84.3.99>
10. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>

11. DCMI Usage Board. (2020). *DCMI metadata terms*. Dublin Core Metadata Initiative. <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>
12. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024). *From local to global: A graph RAG approach to query-focused summarization* (arXiv:2404.16130) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2404.16130>
13. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). Association for Computational Linguistics. <https://aclanthology.org/2024.eacl-demo.16>
14. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey* (arXiv:2312.10997) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2312.10997>
15. Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., & He, Q. (2022). A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8), 3549–3568. <https://doi.org/10.1109/TKDE.2020.3028705>
16. Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, 31(1), 2–24. <https://doi.org/10.1108/EBR-11-2018-0203>
17. Hallo, M., Luján-Mora, S., Maté, A., & Trujillo, J. (2016). Current state of Linked Data in digital libraries. *Journal of Information Science*, 42(2), 117–127. <https://doi.org/10.1177/0165551515594729>
18. Hervieux, S., & Wheatley, A. (2021). Perceptions of artificial intelligence: A survey of academic librarians in Canada and the United States. *The Journal of Academic Librarianship*, 47(1), Article 102270. <https://doi.org/10.1016/j.acalib.2020.102270>
19. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
20. Hogan, A., Blomqvist, E., Cochez, M., d’Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4), Article 71. <https://doi.org/10.1145/3447772>
21. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), Article 42. <https://doi.org/10.1145/3703155>
22. Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D’Souza, J., Kismihók, G., Stocker, M., & Auer, S. (2019). Open Research Knowledge Graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture* (pp. 243–246). Association for Computing Machinery. <https://doi.org/10.1145/3360901.3364435>
23. Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446. <https://doi.org/10.1145/582415.582418>
24. Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 494–514. <https://doi.org/10.1109/TNNLS.2021.3070843>
25. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. <https://doi.org/10.1145/3571730>
26. Kaminskis, M., & Bridge, D. (2016). Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems*, 7(1), Article 2. <https://doi.org/10.1145/2926720>
27. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.550>

28. Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2), 1–224. <https://doi.org/10.1561/15000000012>
29. Lappalainen, Y., & Narayanan, N. (2023). Aisha: A custom AI library chatbot using the ChatGPT API. *Journal of Web Librarianship*, 17(3), 37–58. <https://doi.org/10.1080/19322909.2023.2221477>
30. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates. <https://doi.org/10.48550/arXiv.2005.11401>
31. Library of Congress. (n.d.). *BIBFRAME: Bibliographic Framework Initiative*. <https://www.loc.gov/bibframe/>
32. Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News*, 40(3), 26–29. <https://doi.org/10.1108/LHTN-01-2023-0009>
33. Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5), 570–581. <https://doi.org/10.1002/asi.24750>
34. Miles, A., & Bechhofer, S. (Eds.). (2009). *SKOS Simple Knowledge Organization System reference* (W3C Recommendation). World Wide Web Consortium. <https://www.w3.org/TR/skos-reference/>
35. Oelen, A., Jaradeh, M. Y., & Auer, S. (2024). *ORKG ASK: A neuro-symbolic scholarly search and exploration system* (arXiv:2412.04977) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2412.04977>
36. Padilla, T. (2019). *Responsible operations: Data science, machine learning, and AI in libraries*. OCLC Research. <https://doi.org/10.25333/xk7z-9g97>
37. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, Article n71. <https://doi.org/10.1136/bmj.n71>
38. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7), 3580–3599. <https://doi.org/10.1109/TKDE.2024.3352100>
39. Panda, S., & Kaur, N. (2023). Exploring the viability of ChatGPT as an alternative to traditional chatbot systems in library and information centers. *Library Hi Tech News*, 40(3), 22–25. <https://doi.org/10.1108/LHTN-02-2023-0032>
40. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
41. Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* (arXiv:2205.01833) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2205.01833>
42. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
43. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/15000000019>
44. Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478. <https://doi.org/10.2307/30036540>
45. Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10), 78–85. <https://doi.org/10.1145/2629489>
46. Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, Article 14045. <https://doi.org/10.1038/s41598-023-41032-5>

47. Wang, X., He, X., Cao, Y., Liu, M., & Chua, T.-S. (2019). KGAT: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 950–958). Association for Computing Machinery. <https://doi.org/10.1145/3292500.3330989>.