ADVANCES IN KNOWLEDGE ORGANIZATION
Proceedings of International ISKO Conference - ISSN: 0938-5495 - 2026 - São Paulo

A Knowledge Organization Approach to Smart Retrieval: Combining Controlled Vocabularies and RAG for Semantic Information Retrieval

Juliana Cristhine Schifino Santos Jardim

Federal University of Rio Grande do Sul, Brazil

Rene Faustino Gabriel Junior

Federal University of Rio Grande do Sul, Brazil

Abstract

This study proposes a smart retrieval model integrating Large Language Models with a controlled vocabulary using a Retrieval-Augmented Generation framework. A multilingual thesaurus was developed from ISKO Brazil papers, enabling semantic query expansion and term normalization. The system applies a domain-oriented search strategy in Elasticsearch and uses Large Language Models-based re-ranking to improve relevance. Results show that deterministic parameter settings reduce hallucinations and enhance output reliability. Semantic search guided by controlled vocabulary improves recall without introducing noise, while specific terms and variants increase precision. The findings demonstrate the effectiveness of combining Large Language Models with structured knowledge for improved information retrieval. Keywords: Retrieval-Augmented Generation (RAG). Smart Retrieval. Controlled Vocabulary.

1.0 Introduction

The popularization of generative models, such as ChatGPT in late 2022, has intensified discussions within Knowledge Organization (KO) regarding the use of Generative Artificial Intelligence (GenAI) in Information Retrieval (IR). While Large Language Models (LLMs) demonstrate strong capabilities in interpreting and generating natural language, their application to KO remains limited, particularly due to the lack of domain-specific data in their training corpora.

In this context, Retrieval-Augmented Generation (RAG) architectures gain relevance by enabling LLMs to consult curated external knowledge bases prior to response generation, thereby reducing inaccuracies and hallucinations (Islam 2023). Specialized databases, such as the Information Science Database (BRAPCI), play a key role by aggregating content directly from publishers but continue to face challenges related to incomplete and inconsistent metadata. Missing or inadequate metadata (such as titles, abstracts and authorship) along with low-quality author-assigned keywords, negatively affect retrieval precision.

Given these challenges, this study investigates how LLMs enriched with an external knowledge source can be integrated into knowledge organization and retrieval systems to improve the completeness, specificity, and consistency of bibliographic metadata in Information Science. The general objective is to assess a smart retrieval system implemented through a RAG framework that integrates a specialized thesaurus with LLMs for automatic database indexing. Specifically, the study aims to:

(a) Analyze a RAG-based system by automatically retrieving a sample of ISKO-Brazil articles and comparing its results with those of a non-RAG system; and

(b) evaluate the effectiveness of this approach.

2.0 Revised theoretical framework

The selection and structuring of terms in Knowledge Organization Systems (KOS) are grounded in classical and contemporary principles of concept theory and terminological control. In this study, the process of identifying candidate terms is informed by the notion of scientific relevance proposed by Dahlberg (1978), according to which concepts must correspond to valid and socially recognized units of knowledge within a given domain. This perspective reinforces the epistemological foundation of thesaurus construction, emphasizing that terms should not merely reflect linguistic usage but represent stabilized conceptual structures within scientific discourse.

In addition, the use of term frequency as a criterion for selecting preferred expressions aligns with the contributions of Campos (2001), who highlights the importance of empirical evidence in identifying dominant terminological forms within specialized corpora. Frequency analysis, in this context, supports not only the selection of preferred terms but also the identification of conceptual centrality, which is essential for the hierarchical and ontological organization of knowledge.

The processes of normalization, grouping, and multilingual alignment are further supported by principles of semantic equivalence, as discussed by Herdman et al. (1998). These principles ensure that terms across different languages maintain conceptual consistency, which is particularly relevant in multilingual thesauri and in cross-lingual IR contexts. The distinction between preferred and non-preferred terms, as well as the treatment of synonyms, variants, and acronyms, reflects established practices in controlled vocabulary design, aiming to reduce ambiguity and improve retrieval effectiveness.

Moreover, the adoption of normalization strategies (such as the preference for singular forms and contextual translation) aligns with traditional guidelines in thesaurus construction, contributing to terminological standardization and interoperability. These procedures are essential for enabling semantic integration with computational models, especially in environments that combine structured vocabularies with AI techniques.

Finally, the delimitation of the corpus and the application of controlled search strategies reflect methodological rigor in the construction of domain-specific vocabularies. By prioritizing recent scientific literature and applying linguistic constraints to reduce noise,

the study ensures that the resulting thesaurus is both representative of current discourse and suitable for integration into advanced IR systems.

3.0 Related work

Automatic systems for IR emerged from the need to improve the organization of scientific production, particularly in response to the growing volume of documents anticipated by Luhn (1958). Initially based on statistical methods, indexing evolved to incorporate linguistic and semantic approaches to address ambiguities in natural language (Andreewski and Ruas 1983). Foundational contributions by Lancaster (2004) and Richmond (1974) established automatic indexing as a complementary resource to manual practices, enhancing accuracy, depth, and flexibility in IR.

From the 1980s onward, research focused on systems capable of identifying significant terms through content analysis (Robredo and Ferreira 1980), progressively emphasizing semantic processing and linguistic structures (Robredo 1982; Andreewski and Ruas 1983). Advances in AI, computational linguistics, and knowledge representation in the 1990s and 2000s further shaped information retrieval, with the increasing adoption of Natural Language Processing (NLP), semantic vectors, probabilistic models, and alternative logics to handle uncertainty and contradiction (Silva and Milidiú 1991; Brito 1992; Corrêa and Kobashi 2010). These developments aimed to improve the quality of content representation, a central factor in optimizing core IR metrics such as precision and recall (Manning et al. 2008).

Parallel to these advances, IR evolved from a predominantly academic field into a central mechanism for information access, driven by web search engines, the Semantic Web, and cloud computing infrastructures (Alexander 2003; Viana and Maimone 2025). Traditional models such as TF-IDF and the vector space model laid the foundations for semantic search but remained limited by their lack of contextual understanding.

More recently, the emergence of LLMs based on Transformer architectures has introduced a new paradigm for IR. Trained on large textual corpora, LLMs demonstrate strong capabilities in language generation and interpretation (Géron 2021), but their tendency to generate plausible rather than verifiable content leads to risks of hallucination, particularly in academic contexts (Narayanan and Kapoor 2024). These limitations, combined with the sensitivity of outputs to prompt design and the emergence of unpredictable behaviors at scale, highlight the need for mechanisms that ground and constrain model responses.

RAG addresses these challenges by integrating LLMs with external, curated knowledge bases. By retrieving relevant information through vector search and incorporating it into the prompt, RAG enables the generation of responses that are contextually grounded and factually anchored (Lewis et al. 2020; Gao et al. 2023; Islam 2023). The effectiveness of this approach depends critically on the quality of the retrieval component, including the structure of the knowledge base and the chunking strategy adopted. Therefore, the expertise, as well as the epistemic and mediating role of the human cognition, becomes central to this process.

Within this context, RAG-based systems represent a methodological advancement consistent with the historical trajectory of IR. They provide a more flexible, auditable, and scalable framework (Islam 2023) for enhancing the consistency, specificity, and relevance of metadata generation in KO (Gao et al. 2023). Most importantly, this approach highlights the relevance of information professionals whose expertise enables the production and integration of structured knowledge (particularly in the form of controlled vocabularies) with the contemporary capabilities of LLMs.

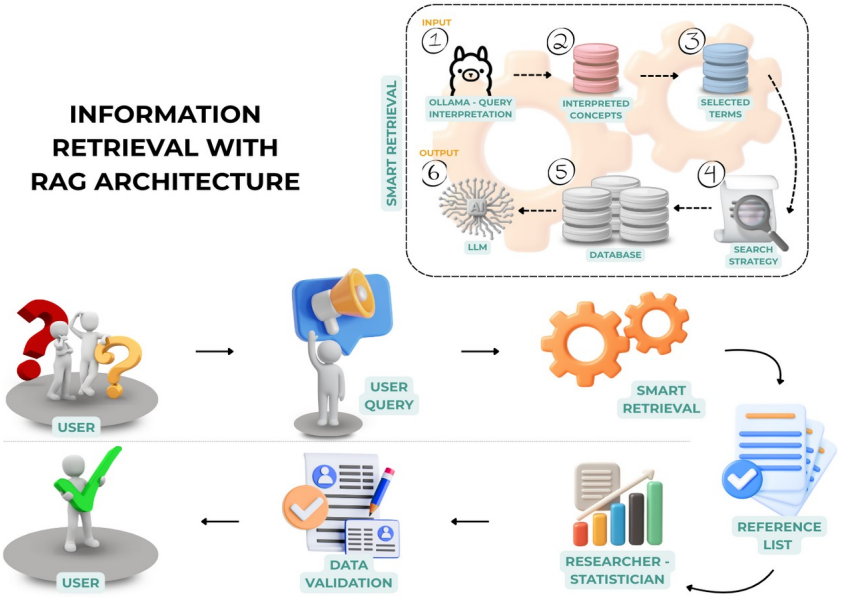

Figure 1. Application of the RAG technique in an Information Retrieval System.

4.0 Methodology

The present study is characterized as applied, as it seeks to develop a model for knowledge organization and retrieval capable of improving practices and outcomes in the field of KO, addressing concrete problems related to representation, indexing, and knowledge retrieval.

A qualitative approach with exploratory and experimental components is adopted, suitable for investigating complex phenomena and for proposing computational and conceptual artifacts. The qualitative approach enables an understanding of knowledge AI–based models, considering the interaction between objective aspects (formal structures, data models, metrics) and subjective aspects (expert decisions, semantic interpretation).

Regarding its objectives, this study is classified as methodological and propositional, as it focuses on the design, development, and validation of a model for knowledge organization and retrieval. This model integrates principles of KO, semantic structures (such as SKOS), and contemporary techniques based on LLMs, with an emphasis on specialized knowledge mediation.

This study aims to propose a model to enhance knowledge retrieval systems by improving both precision and recall. To this end, a six-stage methodological framework is developed, comprising: (i) an architecture integrating Ollama with a controlled vocabulary; (ii) the use

of Ollama to identify relevant terms; (iii) the integration of controlled vocabulary and interpreted concepts; (iv) the application of a query strategy; (v) the execution of queries in a database; and (vi) the retrieval of relevant sources. This integrated approach is referred to as a Smart Retrieval System.

4.1 Architecture to integrate with Ollama

To compose the corpus, a total of 40 papers from ISKO Brazil were selected, comprising 20 from 2019 and 20 from 2025, specifically focusing on AI or closely related topics. Based on these documents, all author-assigned keywords were extracted and used to construct a thesaurus, including semantic relationships (broader and narrower terms) as well as linguistic variations in Portuguese, English, and Spanish.

Using the Thesa API (https://www.ufrgs.br/thesa/web/thesa/25), the controlled vocabulary was integrated into the system and employed within a RAG framework, enabling the enrichment of IR processes through the incorporation of structured domain knowledge.

The resulting corpus comprises 286 concepts, represented by 120 terms in English, 48 in Spanish, and 118 in Portuguese, constituting a multilingual controlled vocabulary. This resource serves as the semantic backbone of the proposed model, supporting tasks such as query expansion, term normalization, and cross-lingual knowledge retrieval, thereby enhancing both precision and recall in the system.

4.2 Ollama for identifying the pertinent terms

To operationalize the use of LLMs for identifying relevant terms, three models were evaluated: LLaMA 3.2, Gemma 3, and Phi-4. The primary selection criterion was the ability to run locally on a server equipped with a GPU with 16 GB of VRAM, ensuring feasibility within a controlled computational environment.

Among the evaluated models, LLaMA 3.2 demonstrated superior performance in terms of semantic interpretation and concept extraction, particularly due to its extended context window of up to 128k tokens, which enabled more consistent analysis of the corpus.

Table 1. Large Language Models Evaluated

Family	Model	Model size	Parameters	Context window	Input
Llama3.2	Llama3.2	2.0GB	3.21B	128k	text
Gemma3	Gemma3	3.3GB	4.3B	128K	text & image
Phi4	Phi4	9.1GB	14.7B	16K	text

To support the model through a RAG approach, a structured prompt (Table 2), originally created in Portuguese, was designed incorporating the controlled vocabulary generated in Thesa. This vocabulary is dynamically injected into each query, allowing the model to ground its responses in domain-specific knowledge and improve terminological consistency.

Additionally, to mitigate hallucinations and increase response reliability, the model was configured with a temperature of 0 and a top_p value of 0.1, prioritizing deterministic and constrained outputs. This configuration ensures greater adherence to the provided vocabulary and enhances the precision of extracted concepts.

Table 2. Prompt

```
prompt = f"""
SYSTEM:
You are a specialized librarian.
Constraints:
• Classify the question using ONLY the most relevant [AUTHORIZED TERMS].
• Identify ONLY the core concepts.
• Do NOT explain your choices.
• Do NOT repeat the extracted terms.
• Do NOT use full sentences.
• Do NOT invent terms.
• Convert plural TERMS to singular.
• Select only specific terms.
• Return ONLY short conceptual terms, separated by commas.
• Do NOT use the term ‘indexing’ (or any variations) in your answer, even if it is relevant.

AUTHORIZED TERMS:
{terms}

USER:
Extract the relevant terms from the question: "{question}"
"""
```

The model interprets the user’s query and processes it through a RAG framework, leveraging a controlled vocabulary to generate a set of semantically relevant terms aligned with the domain-specific knowledge structure.

4.3 Integration of controlled vocabulary and interpreted concepts

Based on the output generated by the Ollama framework, the system identifies relevant concepts and maps them to the Thesa controlled vocabulary, extracting all associated term variants, including synonyms, alternative labels, and orthographic variations. This process ensures terminological normalization and reinforces semantic consistency throughout the retrieval stage.

In addition, the system enables query expansion by incorporating semantically related concepts derived from the thesaurus structure. Through hierarchical and associative relationships, it includes broader, narrower, and related terms, allowing for a more comprehensive representation of the conceptual domain. For example, the concept *“artificial intelligence”* can be expanded to include related concepts such as *Retrieval-Augmented Generation (RAG)* and *Large Language Models (LLMs)*.

This semantic enrichment mechanism enhances both recall and precision by broadening the scope of retrieved documents while maintaining alignment with domain-specific knowledge structures. At the same time, the system organizes the expanded terms into distinct sets according to their respective domains or subdomains. For instance, separate term groups can be generated for *artificial intelligence* and *ethics*, each including their own variants and semantic relationships.

Such domain-oriented structuring enables more refined semantic filtering and supports context-aware query expansion. As a result, the retrieval process can differentiate between distinct conceptual perspectives while preserving their interconnections, leading to more accurate and contextually relevant results.

4.4 The application of a query strategy

Based on the domain-oriented structuring of terms, a search strategy is developed that can be applied across multiple databases, such as Scopus, Web of Science, and OpenAlex. In this study, however, a specific query was designed for implementation in Elasticsearch, allowing the assignment of differentiated weights according to term position and their respective variants.

Within each domain, terms and their variations are combined using the Boolean operator OR, ensuring comprehensive coverage of synonymous and related expressions. Across different domains, the query structure employs the Boolean operator AND, enabling the intersection of conceptual perspectives and refining the retrieval process.

This combined strategy enhances precision by constraining results to documents that address multiple relevant domains, while maintaining satisfactory recall through the inclusion of term variants and semantic expansions.

4.5 The execution of queries in a database and the retrieval of relevant sources

Once the query is formulated, it is executed against the target database, returning a set of documents that typically include metadata fields such as title, abstract, and keywords. This initial retrieval stage is based on the structured query defined from the domain-oriented vocabulary, allowing the system to capture a broad yet semantically guided set of candidate documents.

The retrieved results are then forwarded to the Ollama framework through a RAG pipeline, in which the documents themselves are incorporated as contextual input for the language model. At this stage, the system no longer relies solely on lexical matching but incorporates semantic analysis to assess the relevance of each document in relation to the user's query.

To achieve this, a structured prompt is applied, instructing the model to evaluate and rank the retrieved documents according to their pertinence. The ranking process considers multiple dimensions, including conceptual alignment with the query, contextual coherence between query and document content, and the presence of key domain-specific terms derived from the controlled vocabulary. This step functions as a semantic re-ranking layer, complementing the initial retrieval performed by Elasticsearch.

Additionally, this approach helps mitigate limitations of traditional keyword-based retrieval, such as term mismatch and ambiguity, by leveraging the interpretative capabilities of the language model. As a result, documents that may not share exact lexical matches with the query but are conceptually relevant can be identified and prioritized.

The final output consists of a ranked list of documents, ordered by their estimated relevance to the user's information need. This list is then presented to the user, providing a

more precise and contextually informed retrieval experience, while maintaining a balance between recall and precision through the combined use of structured queries and semantic re-ranking.

5.0 Results

The results demonstrate that model parameter configuration plays a central role in the quality of information retrieval. In particular, the use of temperature = 0 and top_p = 0.1 in the Ollama environment proved essential for reducing the generation of inconsistent or unsupported responses (hallucinations). This more deterministic configuration contributed to greater output stability, ensuring adherence to the controlled vocabulary and improving the reliability of identifying and selecting domain-relevant terms.

Another important aspect concerns the impact of semantic search guided by a controlled vocabulary. The incorporation of domain-specific terms derived from the thesaurus increased the system's retrieval capacity, improving recall without compromising result quality. The system was able to successfully retrieve all 20 documents defined in the test sample without introducing significant noise, demonstrating the effectiveness of the domain-oriented semantic expansion strategy.

Furthermore, it was observed that when users formulate queries using more specific terms, there is a notable increase in precision. In this context, the use of terminological variants (including synonyms, alternative forms, and cross-lingual equivalents) proved essential for overcoming linguistic limitations and domain-specific variations. This approach improved the alignment between user language and system representation, resulting in more precise and semantically consistent retrieval outcomes.

6.0 Conclusion

The proposed model is grounded in recent advances in LLMs, incorporating their capabilities for semantic processing, pattern extraction, and information retrieval support. However, it departs from conventional approaches by shifting the focus from generalist knowledge (predominant in most AI training paradigms) to specialized domain knowledge. This orientation acknowledges that, in contexts such as KOSs, conceptual precision, terminological consistency, and semantic control are central requirements that are not fully addressed by general-purpose models.

In this sense, the model proposes an articulation between the inferential capabilities of LLMs and the intellectual curation performed by domain specialists, fostering more qualified semantic mediation. Rather than replacing the role of the indexer or domain expert, AI technologies are mobilized as supportive instruments, capable of suggesting terms, identifying conceptual relationships, and expanding retrieval possibilities, always under human validation. This perspective aligns with knowledge-driven approaches, as opposed to purely data-driven models.

Furthermore, by privileging specialized knowledge, the model contributes to the development of systems more closely aligned with scientific domain specificities, promoting semantic interoperability, indexing quality, and the reliability of IR processes.

This integration between AI and human expertise points toward a hybrid paradigm in which LLMs act as facilitators of knowledge organization and access, while preserving the epistemological principles that underpin the field.

Acknowledgements and Funding

This research was supported by the Brazilian National Council for Scientific and Technological Development – CNPq, through a Research Productivity Fellowship (Process No. 133119/2025-4), which significantly contributed to the development of this line of investigation.

Statement of AI use

Artificial intelligence tools were used to support translation. The intellectual content, research design, analysis, interpretation of results, and final responsibility for the work remain entirely with the authors.

References

Alexander, Johanna Olson. 2003. “Library Applications.” In Encyclopedia of Information Systems, edited by Hossein Bidgoli, 55–76. Amsterdam: Elsevier. https://doi.org/10.1016/B0-12-227240-4/00104-0.

Andreewski, Alexandre, and Vitoriano Ruas. 1983. “Indexação automática baseada em métodos lingüísticos e estatísticos e sua aplicabilidade à língua portuguesa.” Ciência da Informação 12, no. 1: 61–73. https://brapci.inf.br/v/17283.

Brito, Marcílio de. 1992. “Sistemas de informação em linguagem natural: em busca de uma indexação automática.” Ciência da Informação 21, no. 3: 223–232. https://brapci.inf.br/v/22159.

Campos, M. L. A. 2001. Linguagem documentária: teorias que fundamentam sua elaboração. Niterói: EdUFF.

Correa, Carlos Alberto, and Nair Yumiko Kobashi. 2010. “Estudos de aplicações da lógica paraconsistente em um modelo híbrido de indexação automática.” In Anais do XI ENANCIB: Inovação e inclusão social: questões contemporâneas da informação, 20–28. Rio de Janeiro: IBICT/UFRJ. https://brapci.inf.br/v/177133.

Dahlberg, Ingetraut. 1978. “Teoria do conceito.” Ciência da Informação 7, no. 2: 101–107. https://doi.org/10.18225/ci.inf.v7i2.115.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. “Retrieval-Augmented Generation for Large Language Models: A Survey.” arXiv:2312.10997. https://doi.org/10.48550/arXiv.2312.10997.

Géron, Aurélien. 2021. Mãos à obra: aprendizado de máquina com Scikit-Learn, Keras e TensorFlow: conceitos, ferramentas e técnicas para a construção de sistemas inteligentes. 2nd ed. Rio de Janeiro: Alta Books.

Herdman, M., J. Fox-Rushby, and X. Badia. 1998. “A Model of Equivalence in the Cultural Adaptation of HRQoL Instruments: The Universalist Approach.” Quality of Life Research 7, no. 4: 323–335. https://doi.org/10.1023/A:1024985930536.

Islam, Mohammad Rubyet. 2023. Retrieval-Augmented Generation (RAG): Empowering Large Language Models (LLMs). [S.l.]: Dr. Ray Islam.

Lancaster, F. W. 2004. Indexação e resumos: teoria e prática. 2nd ed. Brasília: Briquet de Lemos Livros. https://bibliotextos.wordpress.com/wp-content/uploads/2014/07/livro-indexac3a7c3a3o-e-resumos-teoria-e-prc3a1tica-lancaster.pdf.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. 2020. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” arXiv:2005.11401. https://arxiv.org/abs/2005.11401.

Luhn, H. P. 1958. “The Automatic Creation of Literature Abstracts.” IBM Journal of Research and Development 2, no. 2: 159–165. https://doi.org/10.1147/RD.22.0159.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Narayanan, Arvind, and Sayash Kapoor. 2024. AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton: Princeton University Press.

Richmond, Phyllis A. 1974. “A Reconsideration of Enumerative Classification for Current Information Needs.” Ciência da Informação 3, no. 1: 5–19.

Robredo, Jaime. 1982. “A indexação automática de textos: o presente já entrou no futuro.” In Estudos avançados em ciência da informação, edited by U. D. Machado, vol. 1, 235–274. Brasília: Associação dos Bibliotecários do Distrito Federal.

Robredo, Jaime, and José Adalberto de Paula Ferreira. 1980. “Conceituação de um programa para indexação automática de textos.” Revista de Biblioteconomia de Brasília 8, no. 2: 254–263. https://periodicos.unb.br/index.php/rbbsb/article/view/29048.

Silva, Wagner Teixeira da, and Ruy Luiz Milidiú. 1991. “Indexação e recuperação da informação com função de crença.” Ciência da Informação 20, no. 2: 155–164. https://brapci.inf.br/v/18776.

Viana, Michelângelo Mazzardo Marques, and Giovana Deliberali Maimone. 2025. “Discovery Systems.” In Encyclopedia of Libraries, Librarianship, and Information Science, 1st ed., vol. 1, 677–688. Cambridge: Academic Press. https://doi.org/10.1016/B978-0-323-95689-5.00072-9.