AKR Advances in Knowledge Representation Journal

Retrieval-Augmented Generation in Information Science: What Do We Need to Know?

Patrícia Nascimento Silva[1] , Daiane Campos Procópio[2] , Danielle Teixeira de Oliveira[3]

Federal University of Minas, Minas Gerais, Brazil

patricians@ufmg.br, campos-daiane@ufmg.br, danielle-oliveira@ufmg.br

Abstract. *Information Science is a field that naturally investigates information, that is, contextualized data, its phenomena and interactions, as well as its impacts on society. With the significant increase in data in digital environments and the rapid evolution of Generative Artificial Intelligence (GAI), these sources have taken on new prominence, as they constitute the raw materials for GAI models. However, identifying how this data and information will be collected, processed and organized for use and retrieval within these models is the central question of this study, which investigates the augmented generation of retrieval, or Retrieval-Augmented Generation (RAG). This research aimed to investigate how the RAG technique has been implemented and used today, in light of information organization and representation, in 2026, through a review of the academic and scientific literature. It was observed that RAG is a technique widely used across various fields; however, in Information Science, it remains relatively uncommon, despite the fact that the field's fundamentals of information organization, representation and retrieval align with the technique, which justifies a fresh perspective on the field to bring it closer to Computing techniques related to Information Retrieval.*

[1] Patrícia Nascimento Silva. Adjunct Professor at the School of Information Science (ECI) of the Federal University of Minas Gerais (UFMG). Professor and Researcher in the Graduate Program in Knowledge Management & Organization (PPGGOC) at ECI/UFMG. Federal University of Minas Gerais. patricians@ufmg.br. https://orcid.org/0000-0002-2405-8536

[2] Ph.D. candidate and Master's degree (2025) in Knowledge Management and Organization from ECI/UFMG. Bachelor's degree in Library Science from UFMG (2013) and in Systems Analysis and Development from PUC Minas (2024). campos-daiane@ufmg.br. Federal University of Minas Gerais. https://orcid.org/0000-0002-9006-191X

[3] Ph.D. candidate and Master's degree (2025) in Knowledge Management and Organization from ECI/UFMG. Coordinator of the Library at the UFMG School of Education. Federal University of Minas Gerais.danielle-oliveira@ufmg.br . https://orcid.org/0000-0002-1958-9113

1. Introduction

Information Science, as a field of knowledge, is a relatively recent science, emerging during a historical period marked by criticism of the failure of the modernity project and its scientific model (Araújo, 2003).

However, its history traces back to "documentation", a term coined by Paul Otlet, in 1903, to describe the process of providing documents for the purpose of information retrieval (Alvares & Araújo Júnior, 2010). During this period, the field was defined as the Science of Documentation by Otlet (Otlet, 1934), until new discussions at various scientific information conferences shifted its trajectory. The mais highlight was the International Conference on Scientific Information, in 1958, at which Robert S. Taylor presented a paper on the glossary of scientific documentation terms, which contributed significantly to defining the field as Information Science (Queiroz & Moura, 2015).

Nevertheless, it was in the 1960s that the field's first concepts and definitions were discussed and developed, and the debate on its origins and theoretical foundations truly began, in an effort to delineate the field and establish interdisciplinary relationships with other fields of knowledge, as well as to envision the role of these emerging professionals (Pinheiro & Loureiro, 1995).

Beginning in the 1960s, information became established as the field's primary object of study, with discussions of classification parameters and theories significant to the field, such as Faceted Analysis Theory (Ranganathan, 1967), Concept Theory (Dahlberg, 1978) and General Theory of Terminology (Wüster, 1931). All of this was accompanied by the technological advancements of the time, such as the emergence of object-oriented programming and relational models, in the 1970s, the first information systems, in the 1980s, and the evolution of cataloging, with the creation of metadata and the *World Wide Web* in 1989 (World Wide Web Foundation, [20–]).

All of this enabled the implementation of expanded and increasingly complex digital environments, with representations involving the semantic web and interoperability, alongside the open access movements of the 2000s, beginning with open research data, followed by open government data and general-purpose and expanded databases, such as Wikidata[4], an open, machine-processable database containing data from various domains.

It was from this period onward, gaining momentum in the 2010s, with mobile devices and their growing popularity, continuing through the 2020s, with more robust *internet* speeds, 5G and phenomena such as Big Data, that significant digital transformations were fueled, accelerated by the COVID-19 pandemic and reaching their peak with Generative Artificial Intelligence (GAI).

In this scenario, an expansion of the field's scope is observed, incorporating data as a central element of research and practice. It is unimaginable to consider all these developments surrounding data without reflecting on the transformations in the field of Information Science, as its fundamentals guide all these new processes and technological revolutions. Thus, in a society experiencing the growing use of emerging technologies

[4] https://www.wikidata.org/.

across a wide range of fields and sectors, the role of Information Science becomes even more comprehensive, requiring constant review and updating of its foundations, objects and practices.

This entire introduction has spanned documentation, information and now data. Therefore, when viewed in a logical context, over the years, the field has increased its granularity, starting from a document that may be associated with the recording of knowledge, moving from information to data.

From this perspective, Retrieval-Augmented Generation (RAG) represents yet another advance in this process of refining information. While traditional Large Language Models (LLMs) respond based on the knowledge learned during their training, RAG allows these responses to be enriched by information retrieved in real time from external sources, such as documents, scientific databases or content available on the internet. By integrating these sources into the response-generation process, the information becomes more up-to-date and contextualized, fostering the construction of more consistent and reliable knowledge (Bourne, 2024). Figure 1 illustrates the logic behind the construction of contextualized knowledge, based on documents and information that will generate new data to be used in RAG.

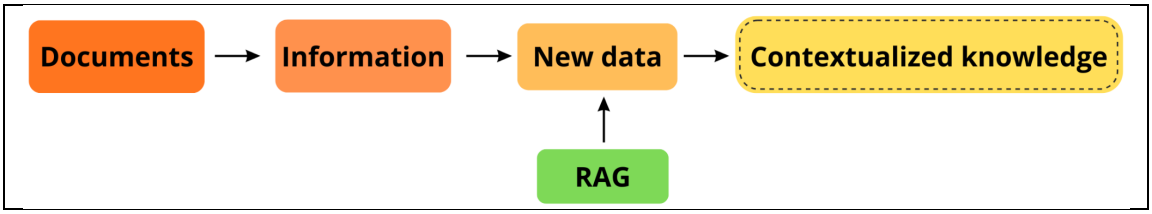

Figure 1 - Representation of RAG[5]

But, beyond this context, the goal of Information Science remains the same: to investigate the origin, collection, organization, storage, retrieval, interpretation, transmission, transformation and use of information, including research on the representation of information in natural and artificial systems, the use of codes for the efficient transmission of messages and the study of processing and techniques applied to computers and their programming systems (Borko, 1968, our translation). As Ferneda (2012) states, identifying, within a document collection, the documents that meet the user's specific information need constitutes the process of information retrieval, which is essential in any informational environment.

Thus, this article promotes the idea that researchers should rethink the opportunities and positioning of the field, which, despite its importance, is not yet widely viewed in symbiotic intersections with new technologies and is losing ground in current discussions.

[5] *Prepared by the authors (2026).*

The application of the RAG technique is recent, dating back to 2020 (Lewis *et al*., 2020), and has grown exponentially. However, there is a hypothesis that this growth in academic and scientific output is more noticeable in technology fields, raising the question: how is RAG being reflected in the context of Information Science? Thus, how the field has been adopting these new techniques is the main question of this study, which conducts a literature review with the aim of investigating the applications and associations of RAG in the field of Information Science.

As pointed out by Cardoso, Ferneda and Souza (2025), there is a gap in the Information Science literature regarding systematic analyses of the applications, potential and ethical challenges of RAG, and there is a need for studies to explore the integration between RAG technologies and Information Science practices, taking into account their social and ethical dimensions. Therefore, this research is justified by its aim to foster a new discussion focused on one of the field's foundations, Information Retrieval (IR), by considering the phases of documentation, information and, now, data as an opportunity to expand upon how Information Science can contribute to informational processes impacted by new technologies.

2. The Evolution of Information Retrieval in Information Science

Information Retrieval (IR) is a broad field that originated in Computer Science, whose central purpose is to enable users to access information that meets their needs and interests (Baeza-Yates & Ribeiro-Neto, 2013). From Saracevic's (1996) perspective, it can also be understood as a technological dimension of Information Science, developed through the convergence of this field with Computer Science.

With the advance of the digital age and the intensification of users' use and appropriation of information, IR has come to incorporate significant contributions from both fields. Thus, the relationship between Computer Science and Information Science, initially marked by conflicts, has become progressively closer, more symbiotic and inseparable, especially since the 1990s.

Based on the studies by Baeza-Yates and Ribeiro-Neto (2013), Luz, Coneglian and Segundo (2019), Procópio (2025), Souza and Rodas (2020), Stock and Stock (2013), data from IBGE (2015) and additional information provided by the authors, Table 1 lists, in chronological order, some of the most significant contributions to the evolution of the IR field, beginning in the 1950s.

Table 1 – Contributions to the field of IR since the 1950s[6]

Year	Researcher(s)	Contribution
1951	Calvin Mooers	Coining of the term "information retrieval".
1955	Allen Kent, Madeline M. Berry, Fred U. Luehrs Jr., and J. W. Perry	Publication of the article *Machine Literature Searching VIII: Operational Criteria for Designing Information Retrieval Systems*, which described precision and recall metrics for Information Retrieval Systems (IRS).
1962	Cyril Cleverdon	Cranfield studies, which evaluated and compared IRs.
1963	Joseph Becker and Robert Hayes	Publication of the first book on information retrieval, titled *Information Storage and Retrieval: Tools, Elements, Theories*.
1968	Gerard Salton	Publication of the first book on information retrieval authored by Gerard Salton, titled *Automatic Information Organization and Retrieval*.
1971	N. Jardine and C. J. van Rijsbergen	Formulation of the clustering hypothesis.
1978	Association for Computing Machinery (ACM)	First ACM conference on information retrieval, ACM SIGIR - *Special Interest Group on Information Retrieval*, in Rochester, NY, which is the leading conference in the field, bringing together researchers from around the world.
1979	C. J. van Rijsbergen	Publication of the book *Information Retrieval*, which discusses probabilistic models.
1983	Gerard Salton and Michael J. McGill	Publication of the book *Introduction to Modern Information Retrieval*, focusing on the vector model.
1989	Tim Berners-Lee	Creation of *the World Wide Web,* which popularized large-scale access to information, driven by search engines such as Yahoo! and AltaVista.
2001	Tim Berners-Lee	Creation of semantic web, which contributed to the development of smarter applications by improving the understanding of data context.
2010s	IBGE	The widespread adoption of mobile devices in Brazil, which expanded access to and streamlined the search for digital content.
2020s	OpenAI	Generative Artificial Intelligence.

[6] *Adapted from Baeza-Yates and Ribeiro-Neto (2013); Luz, Coneglian and Segundo (2019); Procópio (2025); Souza and Rodas (2020); Stock and Stock (2013); and IBGE (2015).*

In the scientific context, the study by Nascimento Silva (2023), which analyzed Brazilian output in IR within the field of Information Science between 2012 and 2021, identified 27 research themes. The studies focused primarily on users and search engines, IR and retrieval models. Starting in 2020, topics related to Artificial Intelligence (AI), chatbots, big data, data science, linked data and information visualization also began to emerge, demonstrating the impact of technological transformations in the field.

3. Methodology

This study is a literature review that is descriptive in nature, as it seeks to highlight the specific characteristics and attributes of a particular group or phenomenon, in this case the RAG technique, and exploratory, as it aims to deepen understanding of the problem under investigation, with a focus on Information Science, thereby providing greater clarity and precision.

Thus, a literature review was conducted to investigate this scenario, which provided the theoretical foundation necessary to understand the current state of knowledge on the topic (Gil, 2017), based on a literature review protocol adapted from Nascimento Silva (2023), detailed in Table 2, which guided the three methodological stages.

Table 2 - Literature review protocol[7]

Literature review protocol	
Criteria	**Description**
General objective	To investigate the applications and associations of RAG in the field of Information Science.
Questions to be addressed	How is the use of RAG being reflected in the context of Information Science? How has Information Science adopted this new information retrieval technique?
Information sources consulted	Databases: ACM Digital Library, BRAPCI, IEE, Scielo, Scopus, Springer Nature and Web of Science.
Eligibility criteria	Language: English, Spanish and Portuguese, to broaden the scope of the publications.
	Availability: documents available in full via open access or through institutional access on the CAPES Journal Portal, for reading and analysis of their content.
	Time frame: no time restrictions.

[7] *Survey data (2026).*

Inclusion and exclusion criteria	Inclusion of documents that address the study's topic.
	Exclusion of duplicate documents.
	Exclusion of documents not written in Portuguese, English or Spanish.
	Documents whose titles and abstracts do not address the research objectives.
	Documents that are not available in full via open access or the CAPES Journal Portal.
Search fields	Title, abstract and keywords.
Search terms	Augmented generation by retrieval and its variations **AND** information science.
	The search terms were used in Portuguese, English and Spanish.
Search strings	(("retrieval-augmented generation" OR "geração aumentada por recuperação" OR "generación potenciada por la recuperación" OR "geração aumentada via recuperação" OR "generación aumentada por recuperación" OR "RAG") AND ("information science" OR "ciência da informação" OR "ciencia de la información"))
Selection procedures for retrieved documents	After applying the eligibility criteria, the titles and abstracts of the retrieved documents will be reviewed to verify the relevance of the content to the overall research objective.
Analysis procedures	A complete reading of the documents to identify elements regarding the use of RAG in the context of Information Science.
Exclusion criteria following document analysis	Works that lack a conceptual, theoretical or methodological approach to the use of RAG in the context of Information Science.
Data management	Google Drive was used to manage research artifacts in a repository. The Parsif.al tool was used to assist in document selection.

The review was conducted using a phased process. In **Phase 1**, the research concepts, terms and synonyms were defined in a structured manner, based on existing research in the field and the authors' prior readings, since no controlled vocabularies were found for the central concept of the research, resulting in a general search string for conducting the searches. This stage was important for maintaining a consistent terminological repertoire aligned with the conceptual axes of the research.

In **Stage 2**, the databases were identified and search strategies were developed and tested, which included a focus on Information Science. Considering the review's objective and it's questions, the selection of information sources was guided by their scientific relevance, scope and academic credibility. Therefore, the following databases were selected: BRAPCI, Scopus, Springer Nature and Web of Science. Searches were

also conducted in the ACM Digital Library, Scielo and IEEE databases; however, no results were found.

It is important to note that this study considered only studies published in traditional, widely recognized databases, even though the topic is frequently addressed in preprint repositories, such as arXiv. While acknowledging the importance of this type of source, it remains sensitive to recent developments. Therefore, we chose to use only peer-reviewed works indexed in databases that adhere to scientific rigor.

The search also included documents in Portuguese, English and Spanish, as these are the languages with which the authors are most familiar. No date restrictions were applied, given that the topic is still recent.

The general search string was then adapted to the syntax of the databases, strictly maintaining the same search strategy to ensure the replicability of the study. The following fields were considered: title, abstract and keywords since expanding the search to all fields available in the database would have resulted in a very large volume of publications, with low precision and high recall. However, the BRAPCI database does not offer a feature to select search fields for advanced searches. Consequently, the search was conducted across all fields, yielding only three results.

The definition of eligibility criteria for the documents retrieved from the databases included journal articles and papers presented at conferences, as well as availability via open access or through institutional access on the Portal of the Coordination for the Improvement of Higher Education Personnel (Capes Portal), which provides federated access to private databases included in this agreement, in accordance with the affiliation of each participating institution in Brazil.

In **Stage 3**, after conducting several tests and validating the searches in each selected database, ensuring they matched the general search string, we proceeded to run the searches, export the results and import them into Parsif.al[8] to support the selection and analysis of the documents. Parsif.al is a tool used to support systematic literature reviews, allowing users to record all selections made, as well as the criteria used and comments for each document.

Thus, within the tool, the protocol created for the review was registered. The results from each database were imported in BIB format, duplicates were automatically identified and eliminated and relevant studies were selected by applying the inclusion and exclusion criteria through a review of titles and abstracts, in order to verify the documents' thematic relevance and their conceptual and methodological alignment with the overall research objective.

Following this initial selection stage, the selected studies were reviewed in full using a detailed spreadsheet designed and filled out manually by the authors. The spreadsheet included metadata and fields defined based on the structure of the documents and information considered relevant to support the analysis, including: identifier, title, document type, language, year of publication, citation, countries or locations addressed, objective, theoretical framework (authors and key themes), methodology (main aspects),

[8] https://parsif.al/.

results, contributions and limitations. This reading process enabled the categorization of the studies according to the theoretical axes of the research, namely: organization, representation and retrieval, facilitating the identification of patterns of convergence, research gaps and predominant theoretical perspectives, thereby supporting the qualitative interpretation and discussion of the findings.

It should also be noted that, given that the article's objective focuses on Information Science, a comparison was conducted among the databases used to highlight this significant difference from the fields of technology. Thus, in addition to the protocol used in the research, this comparison was performed using a search without the Information Science refinement. The results are presented alongside the review, in the next section.

4. Results

This section presents the results of the literature review, including a description of the sample and a discussion of the findings.

4.1 Sample

The searches were conducted considering variations of RAG in English, Portuguese and Spanish in early July 2026, therefore covering all publications up to the first half of 2026, that is, through June 30, 2026. Given the relatively small number of results, care was taken to make some comparisons in order to confirm the study's hypothesis that the technique has been widely used in various fields, but rarely in Information Science.

Consequently, a test was conducted by removing the restriction to Information Science to expand the search, revealing a very significant number of results in Scopus (21,251) and Web of Science (12,140), which demonstrates that the subject area is extensive when the field is not specified. Another modification was also made to increase the number of results by expanding the search to all fields in the databases, not just title, abstract and keywords, while maintaining the restriction to Information Science. It was observed that there were no significant changes, with the exception of Web of Science (116) and Scopus (82,499); however, upon reviewing the first pages of results, it was evident that the precision was very poor, retrieving articles outside the scope of the study.

Thereby, the search defined in the protocol was continued, initially retrieving 205 documents from the four selected databases that returned results. Next, duplicate documents were removed, resulting in only one occurrence per document. Subsequently, a selection was made based on the title, abstract and keywords, resulting in 16 documents. The large number of exclusions is related to the decision not to apply filters to the databases, which resulted in the retrieval of documents that contained some of the search terms but were not related to the scope of the research. The term and architecture *Retrieval-Augmented Generation* (RAG)" were introduced by Lewis et *al*. in 2020, in the paper *"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. However, the searches retrieved publications dating back to 1970. We chose not to set a time frame to verify whether techniques similar to RAG were already in use before 2020, even if

under different names. As a result, the use of the acronym retrieved articles outside the scope of the research, related to other concepts and domains. Nevertheless, we chose to retain it in the search query in order to cover all possible retrieval outcomes.

Before beginning a full reading, a skim read was conducted to obtain an initial overview of the documents. It was observed that four articles were not available via open access, even after authentication through federated access to the Capes Portal; three were written in other languages, such as Mandarin or German; and four others did not fall within the scope of the research in the field of Information Science. Thus, the final sample consisted of five articles, as detailed in Table 3.

Database	Title, abstract and keywords			Skimming	Full reading
	Retrieved	Duplicates	Selected		
Brapci	3	0	3	16	**5**
Scopus	15	0	5		
Springer Nature	183	1	8		
Web of Science	4	0	0		
Total	205	1	16		

Table 3 - Document Selection[9]

4.2 Selected studies

The selected studies demonstrated that research on RAG in Information Science is still in its early stages, with recent publications (2024 - 2026) and the participation of Brazilian researchers (Table 4). A diversity of domains was observed, such as health and cultural heritage, indicating RAG's potential to support different informational processes.

Although the studies are applied in nature, focused on system development and the analysis of use cases, there are discussions that reinforce the need to ground these efforts in the principles of information organization and representation to increase the reliability, contextualization and quality of the responses produced by RAG-based systems.

[9] *Research data (2026).*

Table 4 - Selected studies[10]

ID	Title	Authors	Year	Country (ies)	Objective	Results
1	Brapci as a Tool for Strengthening Open Science: Innovations in the Organization and Dissemination of Information	Rene Faustino Gabriel Junior	2025	Brazil	Describe the innovations developed at Brapci to promote interoperability and the efficient organization of information in digital repositories	Highlights the importance of specialized databases in Information Science in promoting Open Science and democratizing access to scientific information
2	A new "sense of position": from the logic of subject indexing to the geometry of embeddings	João Alberto de Oliveira Lima	2025	Brazil	Propose a conceptual framework demonstrating how text embeddings offer a solution to the problem of "subject indeterminacy"	Demonstrates that *embedding* technology enables a shift from indexing based on rigid human rules to a geometric model driven by data similarity
3	FolkRAG: a retrieval-augmented generation system for cultural heritage materials	Paul Kelly, Jonathan Schild, Amir Jafari	2025	USA	Introduce a proof-of-concept system that uses RAG to integrate and facilitate search across cultural heritage collections	Demonstrates how the integration of metadata from various data sources is used to build a vector database that powers advanced RAG systems, improving access to cultural heritage materials for both novice users and experts, without compromising archival practices
4	Language Models (LLMs) with Retrieval-Augmented Generation (RAG) in Healthcare: Information Science at the Center of	José Eduardo Santarem Segundo	2025	Brazil	To examine technological convergence from a multidisciplinary perspective, in order to present some elements of approach, the state of the	Eight key areas were identified that demonstrate there are many opportunities for research development in the application of LLMs with RAG in the field of health

[10] *Research data (2026).*

	Challenges and Opportunities				art, opportunities and challenges regarding the use and application of LLMs and RAG in the field of healthcare	information
5	Information Retrieval and Generative Artificial Intelligence Using Large Language Models and Retrieval-Augmented Generation	Henrique Leal Tavares *et al.*	2024	Brazil	Present an IR model using a chatbot as an interface and applying GAI, LLM and RAG techniques	Creation of a prototype with an information retrieval environment built using GAI and RAG techniques

4.3 Discussion

Initially, it was observed that there are discrepancies in the definition and characteristics of LLMs among researchers, particularly those from different fields. In the study by Tavares *et al.* (2024, section "Information Retrieval…", para. 6), the definition by Zhao *et al*. (2023) is used to classify LLMs as "Transformer language models containing hundreds of billions (or more) of parameters [...]". However, the Transformer architecture was introduced by Vaswani *et al*. in 2017, with the seminal paper *Attention Is All You Need*. Prior to that, the most widely used architectures were based on Recurrent Neural Networks (RNN) and, later, on Long Short-Term Memory (LSTM) (Santarem Segundo, 2025). Therefore, current LLMs are predominantly based on the Transformer architecture, but the concept is not restricted to this architecture.

There is a divergence of opinions regarding access in the literature. While Santarem Segundo (2025, section "Language models…", para. 4) states, citing Erickson *et al.* (2025), that "easy and cost-effective access to LLMs […] has enabled a generation of developers and researchers to integrate modern LLM-based AI into everyday applications". From another perspective, Barnett *et al*. (2024) report that the use of LLMs is computationally expensive and dependent on high-performance infrastructure, which limits their adoption in contexts with technical or budgetary constraints, even when using approaches such as RAG. In this regard, Carta *et al*. (2024), Siciliani *et al*. (2023) and Raja *et al*. (2024) also discuss how hardware limitations and financial constraints affected the scalability and performance of the systems developed in their research.

However, beyond conceptual issues and perspectives, this research aims to discuss the organization, representation and retrieval of information and knowledge from the perspective of Information Science for the implementation of RAG.

Despite the "ease" provided by the significant number of tools currently available for creating RAG systems integrated with language models, the organization of information and knowledge remains an essential element for ensuring the quality of these applications.

The concept of a knowledge base is closely related to RAG and constitutes a fundamental tool for building high-quality applications. In Information Science, knowledge bases are grounded in the processes of organizing and representing information and knowledge, involving classification systems, taxonomies and various types of metadata used to describe, contextualize and relate data, information and knowledge.

At the same time, curation permeates the entire process of building and maintaining the knowledge base. It is a fundamental activity both for the initial selection of the documents that will comprise the base and for the evaluation and continuous inclusion of new content throughout its life cycle. This process also helps ensure the relevance, reliability, timeliness and thematic appropriateness of the retrieved documents. Santarem Segundo (2025) highlights the importance of curation by noting that, in the RAG, the *prompt* is just as relevant as the curation of the documents that make up the external database.

In this context, the contribution **of information and knowledge organization** to RAG stands out, since these enable the creation of structured and organized databases based on predefined criteria for the selection, arrangement and interrelation of documents, structured by hierarchical or faceted taxonomies. Faceted taxonomies, widely used in today's environments, make it possible to represent the complexity involved in handling a wide variety of information, documents, topics and contexts.

In response to these demands, services designed for knowledge base management have been developed, such as Amazon Bedrock Knowledge, which, in its Titan model, enables faster processing at scale, despite potential challenges with high-volume Application Programming Interface (API) calls (Kelly, Schild & Jafari, 2025). In addition, this service incorporates features related to semantic representation, embeddings and knowledge graphs, making it possible to establish links between content from different data sources. These features are also related to the concepts and applications of ontologies.

When based on ontologies, the RAG technique is called Ontology-Grounded Retrieval-Augmented Generation (OG-RAG). It grounds documents in the domain ontology, converting them into a hypergraph anchored to the ontology and then retrieving a minimal set of hyperedges to reconstruct the context. OG-RAG can improve retrieval performance compared to standard RAG. Ontology-Guided Retrieval-Augmented Generation (ORAG), on the other hand, guides retrieval toward the classification of specific thematic entities based on a label ontology, mitigating hallucinations and improving accuracy in high-performance LLM-based models (Lima, 2026; Sharma, Kumar and Li, 2024; Xiao *et al*., 2024). Cardoso, Ferneda and Souza (2025) point out that ontologies and semantic retrieval can be used to establish deeper connections between concepts and terms used in searches.

This scenario highlights that information scientists can play a significant role in developing applications of this nature, since their training covers fundamentals essential to RAG, although these elements are also investigated by other fields of knowledge and in different application domains.

In the context **of information and knowledge representation**, indexing concepts and processes form the foundation for representing the content of selected and curated documents in machine-processable formats. Defining metadata fields, attributes and schemas in knowledge bases allows documents to be indexed in a contextualized manner, facilitating their identification, relationship-building and subsequent retrieval.

In the study by Kelly, Schild and Jafari (2025), the RAG technique is used to address the limitations arising from fragmented archival metadata and disconnected search tools in extensive archival collections, focusing on the collections of the American Folklife Center, at the Library of Congress. Based on data collection via web scraping and APIs, a vector database was designed to address the lack of standardization and the specialized nature of the metadata present in multiple datasets. The proposal sought to innovatively integrate fundamental principles of library science, such as service, responsibility and authority, with the use of computational tools. To this end, a metadata schema based on a modular approach was developed, with the goal of uniformly

integrating different sources of archival description and enabling their use in a RAG system. With this solution, it was possible to integrate advanced IR methodologies with language generation capabilities, thereby enhancing the interactivity and efficiency of search and navigation within the analyzed folklore collections.

Furthermore, in this context, given the significant volume of data and considering contemporary digital environments with diverse and non-integrated standards, the creation of large knowledge bases, such as so-called commercial data spaces, presents an opportunity for professionals in the field of Information Science. These environments constitute robust services that bring together large datasets from different sources, mostly open data, for use in various domains and, more recently, to feed AI models. This opportunity, however, comes with an essential requirement: data literacy.

As highlighted by Guedes and Nascimento Silva (2025), information professionals need new data-related skills. These professionals must know how to access different sources, whether through data collection via an API or through the handling or analysis of that data. They must also be able to create and manipulate databases, recognize data types and formats, understand data structures and apply the results obtained. These activities often involve creating algorithms and mastering programming languages, but they do not require a programmer to perform them.

Finally, **information retrieval** is the process responsible for locating, selecting and making available the most relevant documents or excerpts to respond to user requests. In RAG, this process can combine semantic search mechanisms, indexing, filters and metadata, so as to consider not only the correspondence between terms but also the context and meaning of the information. The retrieved content is then provided to the language model as supporting elements for generating responses.

According to Cardoso, Ferneda and Souza (2025), retrieval is not limited to a one-off response to a query, but involves the construction of meaning, support for learning and the expansion of users' information literacy. Thus, by combining document retrieval with contextualized text generation, the RAG technique expands the possibilities for meeting users' informational needs.

The RAG technique represents a significant advance in natural language processing by combining LLMs with targeted information retrieval capabilities from external data sources stored in vector databases, enabling them to overcome limitations such as hallucinations and static training data. The core architecture comprises a retriever that locates relevant information from the vector database and a generator that incorporates this information to produce accurate, context-rich responses Lewis *et al.* (2020); Veturi *et al.* (2024); Guu *et al.* (2020); Karpukhin *et al.* (2020); Izacard and Grave (2020).

Thus, in RAG, the quality of retrieval also depends on the organization, representation and curation involved in building this vector database, reaffirming the contribution of the fundamentals of Information Science to the development of more contextualized, accurate and reliable RAG applications.

5. Concluding Remarks

Given the technological and phenomenological evolution of recent years, in parallel with the field of Information Science, it is evident that the field maintains its purpose, but it is undeniable that it must keep pace with changes, especially technological ones, that alter the way information is produced and used.

With IAG, which gained popularity starting in 2022 and was initially more prevalent in technology fields, various fields of knowledge have come to have access to tools and related techniques to support different activities and processes. Among these, the RAG technique stands out, as it uses external sources to optimize, enhance, and contextualize scenarios. Within this context, the research question for this study emerged, focusing on how Information Science is integrated into these processes. Thus, the study aimed to investigate the applications and associations of RAG in the field of Information Science, based on a literature review conducted in 2026. It should be noted that this study did not use preprint databases, as they do not contain peer-reviewed publications, despite the significant volume of work on the subject.

As a result, 205 documents were initially selected and, after applying the inclusion and exclusion criteria and conducting a full review, only five documents were included in the final sample. This is a very small number, but it was validated in the databases, and this limitation is directly related to the scope of Information Science.

When analyzing RAG technology from the perspectives of organization, representation and retrieval, it is evident that Information Science is present, whether in the creation of taxonomies and classifications for organization, in indexing and metadata assignment for representation, or in natural language and semantic aspects for retrieval. Furthermore, the factor that could expand the role of information professionals in this field is data literacy, which involves the ability to understand and analyze data structures, as well as to use technological resources, such as databases and programming languages, to manipulate them.

Given this context, information professionals need to reinvent themselves and adapt to changes in order to continue applying the fundamentals of the field in different contexts, just as occurs in other areas of knowledge. Moreover, it is important to offer complementary training alongside undergraduate and graduate programs that encourage professionals to seek new knowledge and embrace challenges, stepping out of their comfort zones to reach new heights.

For future work, it is suggested that this technique be applied, whether through established tools or by customizing models, to specific cases in Information Science, in order to foster new applications in the field.

Acknowledgments

To the National Council for Scientific and Technological Development (CNPq) for its support of this research, grant number 303721/2025-1.

To the Coordination for the Improvement of Higher Education Personnel (CAPES) for the financial support provided.

6. References

Alvares, L. & Araújo Júnior, R. H. de. (2010). Historical Milestones in the Science of: A Brief Chronology of Pioneers, Classic Works, and . *Transinformação*, 22(3), 195–205. https://www.redalyc.org/pdf/3843/384334885001.pdf

Araújo, C. A. Á. (2003). Information science as a social science. *Ciência da Informação*, 32(3), 21–27. https://doi.org/10.1590/S0100-19652003000300003

Baeza-Yates, R. & Ribeiro-Neto, B. (2013). *Information Retrieval*: Concepts and Search Engine Technology (2nd ed.). Bookman.

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven Failure Points When Engineering a Retrieval-Augmented Generation System. In *Proceedings of the 3rd International Conference on AI Engineering – Software Engineering for AI* (pp. 194–199). ACM. https://doi.org/10.1145/3644815.3644945

Borko, H. (1968). Information science: What is it? *American Documentation*, 19(1), 3–5. https://doi.org/10.1002/asi.5090190103

Bourne, K. (2024). *Unlocking Data with Generative AI and RAG*. Birmingham: Packt.

Cardoso, F., Ferneda, E., & Souza, C. B. dos S. de. (2025). The Contribution of Retrieval-Augmented Generation to Information Science: Innovations, applications, and ethical implications. In XI EDICIC Ibérico: Education and Research in Information Science: Opportunities and Challenges. https://doi.org/10.34630/xiedicic.vi.6699

Carta, S., Giuliani, A., Manca, M. M., Piano, L., Podda, A. S., Pompianu, L., & Tiddia, S. G. (2024). A zero-shot strategy for knowledge graph engineering using GPT-3.5. *Procedia Computer Science, 246*, 2235–2243. https://doi.org/10.1016/j.procs.2024.09.573

Dahlberg, I. Concept Theory. (1978). *Information Science*, 7(2), 101–107. https://upload.wikimedia.org/wikipedia/commons/6/66/Teoria_do_conceito_%28Cionline_115%29.pdf

Erickson, J. S., Santos, H., Pinheiro, V., McCusker, J. P., & McGuinness, D. L. (2025). LLM experimentation through knowledge graphs: Towards improved management, repeatability, and verification. *Journal of Web Semantics*, 85, Article 100853. https://doi.org/10.1016/j.websem.2024.100853

Ferneda, E. (2012). *Introduction to computational models of information retrieval*. Ciência Moderna.

Gabriel Junior, R. F. (2025). Brapci as a tool for strengthening Open Science: innovations in the organization and dissemination of information. *Revista Ciência Aberta Lusófona*, 1, 289–296. https://doi.org/10.82524/recal.2025.5

Gil, A. C. (2017). How to Design Research Projects (6th ed.). Atlas.

Guedes, S. F. M. de S. & Nascimento Silva, P. (2025). Data literacy for information professionals: a proposed app for the prevention and management of environmental disasters in Belo Horizonte. *ACB Journal*, 30(1), 1–28. https://revista.acbsc.org.br/racb/article/view/2216

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M-W. (2020). Retrieval-augmented language model pre-training. In: International Conference on Machine Learning [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2002.08909

Brazilian Institute of Geography and Statistics (IBGE). (2015). *Pnad 2013*: Mobile internet is used in more than half of households that access the internet. IBGE News Agency. https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/9840-pnad-2013-internet-pelo-celular-e-utilizada-em-mais-da-metade-dos-domicilios-que-acessam-a-rede

Izacard, G. & Grave, E. (2020) Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering [Preprint]. *Arxiv*. https://doi.org/10.48550/arXiv.2007.01282

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yin, W-T. (2020). Dense passage retrieval for open-domain question answering [Preprint]. Arxiv. https://doi.org/10.48550/arXiv.2004.04906

Kelly, P., Schild, J., & Jafari, A. (2025). FolkRAG: A retrieval-augmented generation system for cultural heritage materials. *Neural Computing and Applications, 37*, 20281–20297. https://doi.org/10.1007/s00521-025-11455-4

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks [Preprint]. *Arxiv*. https://doi.org/10.48550/arXiv.2005.11401

Lima, J. A. de. O. (2025). A new "sense of position": From the logic of subject indexing to the geometry of embeddings. *Journal of Documentation, 82*(1), 136–153. https://doi.org/10.1108/JD-07-2025-0204

Luz, L. P., Coneglian, C. S., & Segundo, J. E. S. (2019). Semantic Web Technologies for Information Retrieval in Wikidata. *Digital Journal of Library and Information Science*, 17, e019003. https://doi.org/10.20396/rdbci.v17i0.8651791

Nascimento Silva, P. (2023). Information Retrieval in Information Science: Brazilian Academic and Scientific Output (2012–2021). *Transinformação*, 35, e237336, 1–17. https://doi.org/10.1590/2318-0889202335e237336

Otlet, P. (1934). *Traité de documentation: Le livre sur le livre, théorie et pratique*. Editions Mundaneum.

Pinheiro, L. V. R. & Loureiro, J. M. M. (1995). Outlines and boundaries of information science. *Ciência da Informação, 24*(1). https://revista.ibict.br/ciinf/article/view/609

Procópio, D. C. (2025). *Large language models for information retrieval in digitized historical collections* [Master's thesis, Federal University of Minas Gerais]. Institutional Repository of the Federal University of Minas Gerais. https://hdl.handle.net/1843/3339

Queiroz, D. G. de C. & Moura, A. M. M. de. (2015). Information Science: history, concepts, and characteristics. *Em Questão*, 21(3), 26–42. https://seer.ufrgs.br/index.php/EmQuestao/article/view/57516

Raja, H., Munawar, A., Mylonas, N., Delsoz, M., Madadi, Y., Elahi, M., Hassan, A., Abu Serhan, H., Inam, O., Hernandez, L., Chen, H., Tran, S., Munir, W., Abd-

Alrazaq, A., & Yousefi, S. (2024). Automated category and trend analysis of scientific articles on ophthalmology using large language models: Development and usability study. *JMIR Formative Research, 8*, e52462. https://doi.org/10.2196/52462

Ranganathan, S. R. (1967). *Prolegomena to Library Classification* (3rd ed.). London: Asia Publishing House.

Santarem Segundo, J. E. (2025). Language models (LLMs) with retrieval-augmented generation (RAG) in healthcare: Information Science at the center of challenges and opportunities. In *Proceedings of the 25th National Conference on Research and Graduate Studies in Information Science (ENANCIB)*, 5(1). https://enancib.ancib.org/enancib/pt_BR/article/view/1195

Saracevic, T. (1996). Information Science: Origins, Evolution, and Relationships. *Perspectives in Information Science*, 1(1), 41–62. https://periodicos.ufmg.br/index.php/pci/article/view/22308

Sharma, K., Kumar, P., & Li, Y. (2024). OG-RAG: Ontology-grounded retrieval-augmented generation for large language models [Preprint]. *Arxiv*. https://arxiv.org/abs/2412.15235

Siciliani, L., Taccardi, V., Basile, P., Di Ciano, M., & Lops, P. (2023). AI-based decision support system for public procurement. *Information Systems, 119*, Article 102284. https://doi.org/10.1016/j.is.2023.102284

Souza, R. A. S. & Rodas, C. M. (2020). Information Retrieval on Mobile Devices. *Biblos*: Journal of the Institute of Humanities and Information Sciences, 34(2), 147–166. https://doi.org/10.14295/biblos.v34i2.11840

Stock, W. G. & Stock, M. (2013). *Handbook of information science*. De Gruyter.

Tavares, H. L., Coneglian, C. S., Torino, E., Vidotti, S. A. B. G., & Santarem Segundo, J. E. (2024). Information retrieval and generative artificial intelligence with large language models and retrieval-augmented generation. In *Proceedings of the XXIV National National Conference on Research and Graduate Studies in Information Science (ENANCIB),* 4(1). https://enancib.ancib.org/enancib/pt_BR/article/view/1801

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. In I. Guyon *et al.* (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Veturi, S., Vaichal, S., Jagadheesh, R. L., Tripto, N. I., & Yan, N. (2024). RAG-Based Question-Answering for Contextual Response Prediction System [Preprint]. *Arxiv*. https://doi.org/10.48550/arXiv.2409.03708

Xiao, J., Ding, L., Barry, J., Elkaref, M., De Mel, G., & Han, J. (2024). Ontology-Guided Retrieval-Augmented Generation for Theme-Specific Entity Typing. In *Proceedings of the First Conference on Language Modeling (COLM 2024)*. https://openreview.net/pdf?id=cKBmZ2PZ6c

World Wide Web Foundation. (n.d.). *History of the Web*. https://webfoundation.org/about/vision/history-of-the-web/

Wüster, E. (1931). *International Language Standardization in Technology, Especially in Electrical Engineering*: National Language Standardization and Its Generalization. VDI-Verlag.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023). A survey of large language models [Preprint]. *Arxiv*. https://doi.org/10.48550/arXiv.2303.18223