Research Proposal for Dissertation:

This research proposal is for a manuscript-based dissertation.

Introduction

In the last few years, artificial intelligence (AI) has been rapidly deployed in academic libraries as an aid for discovery (Concensus), tool for systematic review (Han et al., 2024), identifying citation contexts (Scite.ai), or to blatantly promote publisher-owned works (Scopus AI, Web of Science Research Assistant, ProQuest One, and PrimoVE Research Assistant). Large language models (LLMs) are excel at generating summaries or answering questions and within academic environments, there is a need for verifiable sources. Retrieval-augmented generation (RAG) architectures solve this problem by using external knowledge sources. As an open source of data, bibliographic metadata, as an open source of information, has potential to serve as a source of knowledge for these systems.

Concurrent to this research, discussions and task groups are working on what community enrichment of metadata looks like and how it might scale to be introduced in large bibliographic databases. COMET, a task group assembled by the California Digital Library[1], is currently working on enrichment projects to improve funding metadata, classification, and affiliation parsing using community members beyond those at bibliographic databases or publishers. OpenAlex is currently having discussions about metadata enrichment, but they are interested in how they may improve metadata by aggregating it from libraries where considerable work is done to improve discoverability in local discovery systems (P. Riddle, personal communication, August 2025). Registration agencies such as Crossref and DataCite are already working towards integrating such data, with Crossref's development roadmap anticipating the need for this (Hendricks, 2025), and DataCite posting a request for comment from the community on their proposed schema (datacite, 2025).

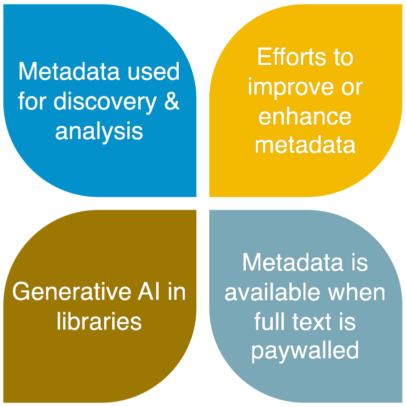

Figure 1: Four directions shaping this research proposal.

[1] https://www.cometadata.org/about

Three areas should be explored to understand bibliographic metadata as an external knowledge source for grounding RAG applications: metadata quality issues that affect RAG configuration decisions, how metadata sources compare for these elements, and an example of what this looks like in action with a case study.

Problem statement

Metadata of scholarly works, specifically the title and abstract, can be used as an external source of knowledge, not only just a representation of the full text. While previously used for discovery or analysis, these open resources can be used to ground RAG applications. There are two unknowns with this approach. One concern is abstract and title content. While abstract and title elements have been investigated for coverage and availability, a detailed analysis to characterize their content from the perspective of a natural language processing (NLP) task has not been performed. The second concern relates to the source of metadata as abstracts are not routinely deposited as part of the metadata record assigned to a DOI, but they may be added as part of an aggregation process in other databases thus introducing unintended content containing problematic characteristics or errors.

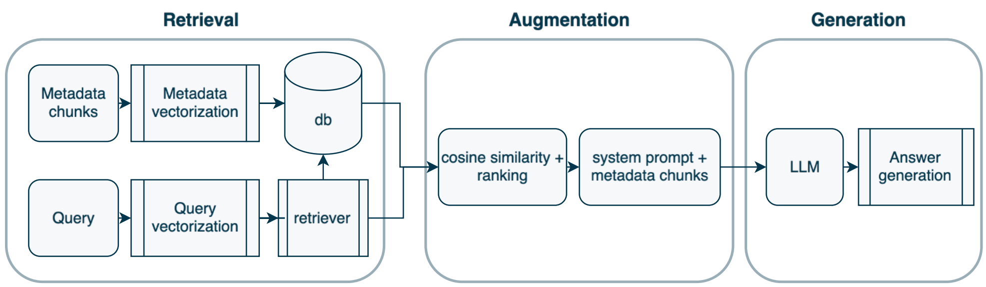

Fig 1: Diagram of a basic RAG architecture.

RAG systems leverage LLM's for response generation but augment the internal knowledge of the LLM with external knowledge addressing issues of discipline specific, proprietary, more recent, or other sources of knowledge that were out of scope of the training data for the LLM. Metadata from scholarly publications may be used as a document source instead of the full text that the metadata represents to ground LLM responses with truthful and verifiable answers. RAG applications have several advantages in that their grounding data can be dynamically updated, be private or confidential, or be more up to date than the training data used for the LLM. Additionally, RAG applications can be quickly and easily adapted to provide answers for different users by changing the grounding documents. One of the attractive benefits of RAG applications for scholarly works is being able to quickly learn about a topic through question-and-answer (Q&A) interaction and this is the scenario in which publicly available bibliographic metadata can be an external knowledge source for grounding RAG applications.

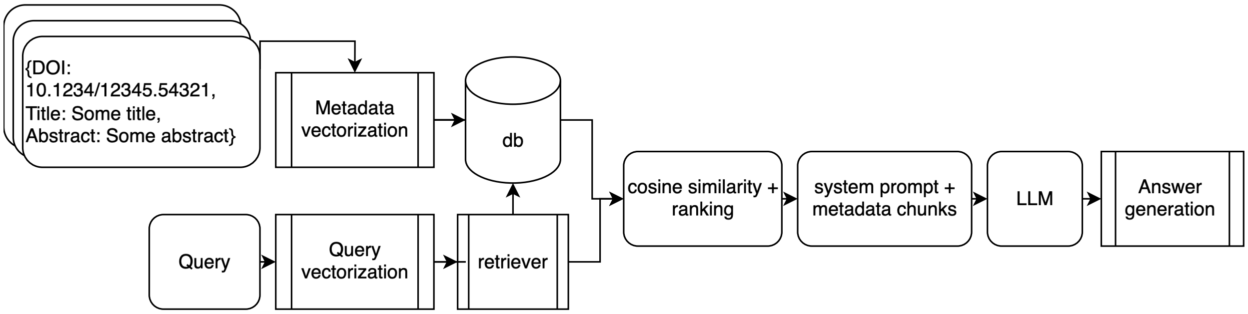

Fig 2: Where metadata elements of title, abstract, and DOI fit in the RAG pipeline.

However, the first issue of metadata content in this scenario particularly affects the abstract element of the metadata as it is the primary source of knowledge for the RAG application. One abstract may not be enough to answer a question, but many abstracts may contain enough information to provide answers to user queries. This is based on preliminary work conducted in the Summer of 2024 by (Saiyed et al., 2025) and similar prior research (Kang & Kim, 2023). The information needed for a RAG depends on the types of questions being asked, so for this context, I'll focus on a scenario in which the descriptive metadata would provide answers to discipline specific questions that would be expected for someone learning about a topic. One gap is a lack of characterization of the data contained within elements such as <abstract> and <title> from the perspective of what may be a challenge for a natural language processing (NLP) task (gap 1). Prior research has found that typos (Cho et al., 2024), noisiness due to irrelevant information (Chen et al., 2024; F. Shi et al., 2023), contradicting (Wan et al., 2024), and counterfactual (Longpre et al., 2022; Marjanović et al., 2024) information negatively impact LLMs whether they are used for embedding or generation. Understanding the breadth of characteristics of these elements, and their availability has implications for text cleaning, tokenization, out-of-vocabulary (OOV) terms, stop word and special character handling, chunk file preparation, and chunk strategies such as truncation, which can be important for embedding model decisions. Additionally, it may be beneficial for the research community to be introduced to the schema definitions as these seem to be overlooked in the literature.

The second issue of metadata source and its aggregation and transformation from its origin point at a registration agency (RA) to aggregators is not well known at a detailed level of element content or attributes. Past studies have compared sources such as Crossref and OpenAlex for their coverage (Culbert et al., 2024; Delgado-Quirós & Ortega, 2024; Schares, 2024; Scheidsteger & Haunschild, 2023; Zhang et al., 2024) and change of element values during ingest one schema to another (Haupka et al., 2024), but what is needed in this context is an understanding how the textual content may change, such as token counts in abstracts as this may affect downstream decisions of clean, truncation, or language alignment for embedding model selection. As aggregators obtain data from multiple sources in addition to adding their own data derived from web scraping or from full text, more information may be available than from the original source (from the RA), or the text may be altered, or language attributes may be lost. Knowing these differences would affect RAG configuration decisions, (such as text cleaning steps or selection of appropriate embedding models) and inform our knowledge of limitations when using the metadata as an external source. At this time, selecting a source for metadata can be based on past studies for coverage or availability, but there is a lack of supporting evidence about the characteristics and errors in metadata from sources (Crossref or OpenAlex) for the purpose of using the metadata as a document source (gap 2).

To illustrate what satisfying these two gaps looks like in action, a case study using a created corpus as a source for a RAG Q&A application would be useful to show how title and abstract content characteristics and errors affect RAG configuration decisions, specifically, how errors or characteristics identified in Part 1 and associated with source decisions from Part 2 affect retrieval and generation outcomes. A corpus of selected works will be duplicated and using one as a control, modify the content of the second using errors and characteristics. Comparing retriever and generation outcomes may illustrate how metadata errors/characteristics impact configuration decisions for the retriever model, ranking, and generation model. A case study could contribute to the literature about what metadata elements may be used as a document source, how metadata characteristics may be affected by choice of where its harvested and illustrate how metadata quality affects outcomes for a given set of queries.

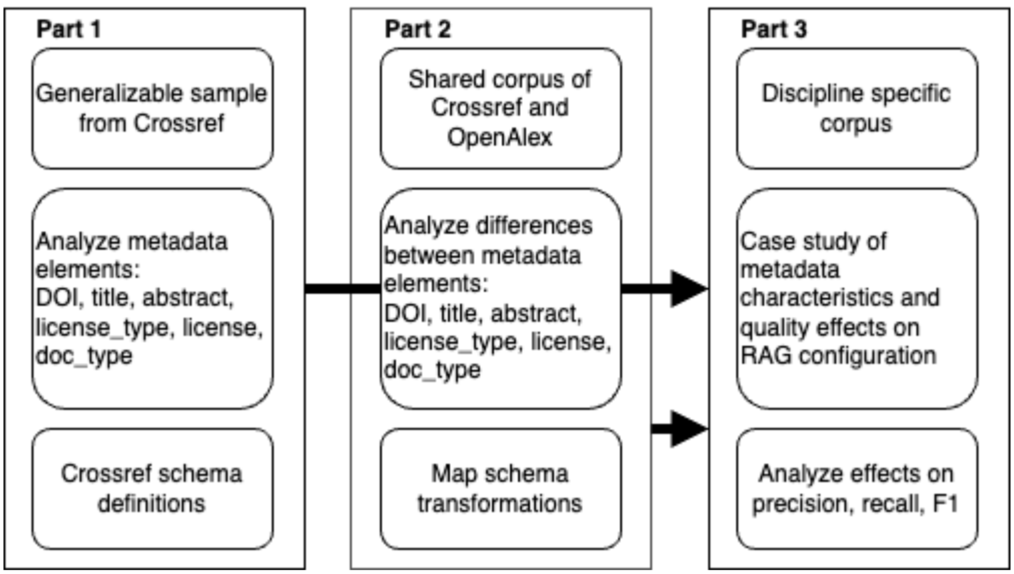

Fig 3: research framework

Theoretical position

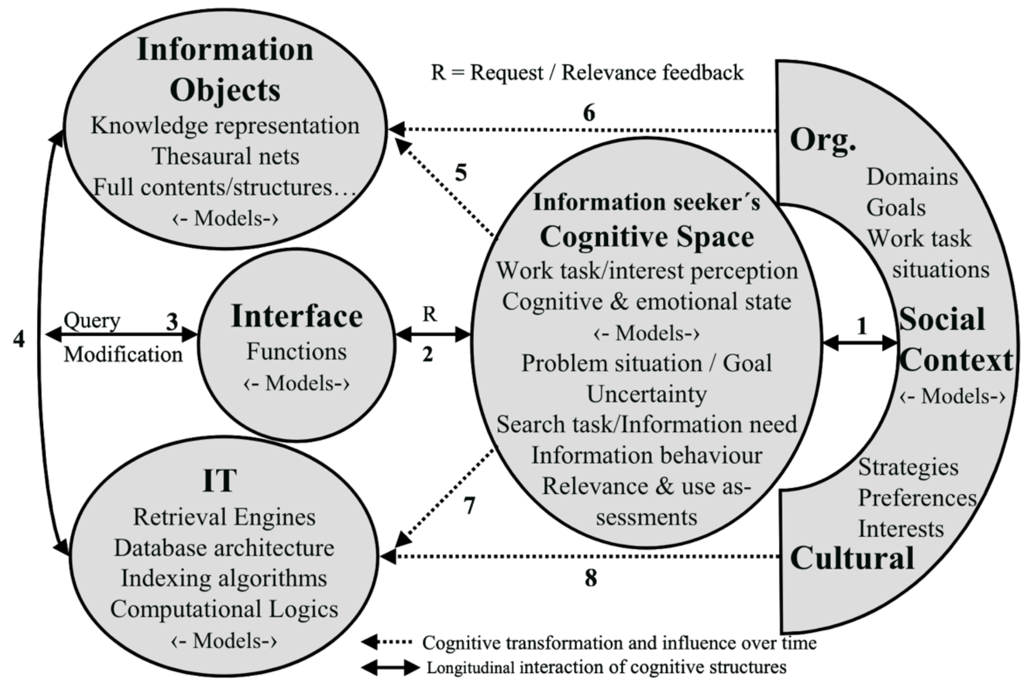

Figure 2: From Ingwersen & Järvelin (2005, p. 271), the interactive information seeking, retrieval and behavioural processes which they describe as "a generalized model of any participating cognitive actor in context".

Ingwersen & Järvelin's cognitive information search and retrieval, (IS&R) model (Ingwersen, 1996; Ingwersen & Järvelin, 2005) extended what was the traditional laboratory model of information retrieval (IR) research to include the cognitive world of the user/information seeker while simultaneously acknowledging the impact that social, cultural, organization, and technological influences upon their cognitive state. "…[B]oth information seeking and information retrieval fall short regarding goals of a) theoretically understanding the phenomena of information access, b) empirically describing (predicting) these phenomena, and c) supporting the development of technology covering …tools, systems, and social practices" (Ingwersen & Järvelin, 2005, p. 377).

The IS&R model has five main components, with data structures representing the cognitive structures of the actors involved over time. Contexts are historical or nested within the components and interactions between components affect the actor's perception of the situation. The components offer a way of illustrating the perceptions and context of the actor in their world and identifying limits on influencing links. The model also addresses not only information seeking, but also information behaviour, one nested in the other as discussed as nested contexts by Wilson (1999). The cognitive IS&R framework's components can be used to explore detailed elements within each component to "support a more structured and detailed way of investigating central issues in IS&R" (Ingwersen & Järvelin, 2005, p. 307).

A cognitive model is the representation of the actor's state of knowledge and can include their "cognition, expectations, emotions, intentionality, experiences, imagination, intuition, values, and environment” (Ingwersen & Järvelin, 2005, p. 382). Underlying the cognitive model is the epistemological perspective that the cognitive model is a constructed, subjective thing including emotions that are determined both by the individual and their context, including but not limited to experiences, education, social, organizational, and cultural influences, and systemic factors.

There are five main components of the cognitive model of IS&R. All of these are involved in information processes at any given time and over time. The arrows denote the direction of mutual interaction involving and affecting cognitive structures. Likewise, there are arrows where influence only goes in one direction, such as in the one labeled 6 above, where social, cultural, and organizational influence affects information objects. This is not a process model but can be used to understand processes that exist because of the interactions between components of the model. At any time, all of the components are in play as they shape the context of the user. However, not all components need, or should be investigated simultaneously.

Each component of the model expands into a nine-dimensional design cube, (represented as a table below). As the authors state, “One may use the 9 dimensions as a checklist for what should be taken into account when designing an investigation” (Ingwersen & Järvelin, 2005, p. 360). Each study may focus completely on one component and one dimension for its descriptive or experimental study, but it is possible to use multiple components and their dimensions to account for controlled variables or assumptions. In Table 1 below, the top row indicates the five components with some components containing more than 1 dimension. Below each dimensional heading are the individual dimensions which can be assigned as independent or dependent in the case of experimental studies, or observed variables in the case of descriptive studies, along with controlled variables.

*Table 1: Design cube of nine dimensions with main components at the top. The top row contains the component names followed by the second row where some components have more than one dimension. Type in red and marked with * indicates changes I have made to adapt the model to working with metadata and RAG applications.*

Organizational Task Component		Actor Component			Document Component	Algorithmic Component		Access and Interaction Component
Natural work tasks and organizational	**Natural search tasks (ST)**	**Actor**	**Perceived work tasks**	**Perceived search tasks**	**Document and source**	**IR methods and NLP models***	**IR interfaces**	**Access and Interaction**
WT structure	ST structure	Domain knowledge	Perceived WT structure	Perceived information need content	Document structure	Exact match models	Domain mode attributes	Interaction duration
WT strategies and practices	ST strategies and practices	IS&R knowledge	Perceived WT strategies and practices	Perceived ST structure and type	Document types	Best match models	System model features	Actors or components
WT granularity, size, and complexity	ST granularity, size, and complexity	Experience on work task	Perceived WT granularity, size, complexity	Perceived ST strategies and practices	Information type in document	Degree of document structure and context used (chunking)*	User model features	Kind of interaction and access

WT dependencies	ST dependencies	Experience on search task	Perceived WT dependencies	Perceived ST specificity and complexity	Communication function	Use of NLP for document vectorization*	System model adaptation	Strategies and tactics
WT requirements	ST requirements	Stage in WT execution	Perceived WT requirements	Perceived ST dependencies	Temporal aspects	Document metadata representation	User model building	Purpose of human communication
WT domain and context	ST domain and context	Perception of socio-organizational context	Perceived WT domain and context	Perceived ST stability	Document sign language	Use of weights in document vectorization*	Request model builder	Purpose of system communication
		Sources of difficulty		Perceived ST domain and context	Layout and style	Degree of required structure and context used	User retrieval strategy	Interaction mode
		Motivation and emotional state			Document metadata*	Use of NLP for request vectorization*	Response representation *	Least effort factors
					Document content	Request metadata representation *	Feedback generation	
					Contextual hyperlink structure	Use of weights in query vectorization*	Mapping ST history	
					Data source*	Response generation*	Explanation features	
							Transmission of messages	
							Scheduler	

From the perspective of this proposal, the Algorithmic component's dimension has been renamed to 'IR methods and NLP models' and dimensions have been indicated in red for those that have been changed to update the design cube to the needs of a study on RAG. This includes aspects of LLMs for embedding and generation and moving 'Response generation' from the 'IR interfaces' dimension to that of 'IR methods and NLP models'. These changes are not exhaustive for technology present in 2025 but I only include changes in the Document and Algorithmic components that are relevant to this research.

Initially from Ingwersen's doctoral dissertation and building on the cognitive model of communication (Belkin, 1984), the IS&R model introduced the concept of interaction in IR involving more than just the user, including an intermediary and the system (Ingwersen, 1992). Saracevic further stratified this into three levels including the user, the system, and a cognitive level (Saracevic, 1996). While Inwersen & Jarvelin and Saracevic further published works on the complexities of relevance judgements, the model from 2005, (sometimes called the integrative model of IS&R (Savolainen, 2018)) is being used for its practical usefulness in identifying and acknowledging dimensions that situate a research area within a system of components and their influences. Similar adaptation of the IS&R model can be seen in Pawlick-Potts (2022) who made modifications to the actor, IT, and information objects components.

No studies have investigated the title and abstract to understand what characteristics exist in the content and how this may affect RAG performance when using the title and abstract as the external knowledge source. Metadata quality problems have been defined by Yasser (2011) and at this time, their classification still seems appropriate as the basis for labelling errors: “Incorrect Values, Incorrect Elements, Missing Information, Information Loss, and Inconsistent Value Representation” (Yasser, 2011, p. 51). Prior research has investigated Crossref, OpenAlex and other databases for metadata coverage or availability (Alperin et al., 2024; Eck & Waltman, 2022), completeness (Delgado-Quirós & Ortega, 2024), error types found in metadata, and investigated with a focus on particular metadata elements such as funder (Kramer & de Jonge, 2022; Mugabushaka et al., 2022), license (Schlosser, 2016), references (Culbert et al., 2024), DOI errors (Cioffi et al., 2022), document types (Haupka et al., 2024; Mongeon, Hare, Krause, et al., 2025), language (Céspedes et al., 2024; Mongeon, Hare, Riddle, et al., 2025; J. Shi et al., 2025), journals (Mongeon & Paul-Hus, 2016), institutions (Zhang et al., 2024), abstracts (Delgado-Quirós & Ortega, 2024; Färber et al., 2022; Kramer & de Jonge, 2022) and varying combinations. While the Eck & Waltman, (2022) study provided evidence to support selecting Crossref or OpenAlex as a source for analysis based on availability of abstracts, there is a lack of analysis on the title and abstract content to understand what might affect NLP processes.

No study has been conducted that compares the title and abstract elements from multiple sources to understand how aggregations and transformations affect the content. Some studies have investigated the translation of journal types or topic areas (Santos et al., 2023), have identified how transformations can introduce information loss (Yasser, 2011), or introduce problems that did not exist in the original metadata deposit (Delgado-Quirós & Ortega, 2024). Despite this, Eck and Waltman show that certain elements in Crossref metadata have improved over time, especially for journal articles (Eck and Waltman, 2022). While ”Crossref is becoming an increasingly interesting data source for bibliometric analyses” it (and possibly OpenAlex) may also be an interesting external knowledge source for RAG applications.

The case study serves as a novel application of metadata as a source document in a RAG application, particularly for the LIS community. Building on the work of (Saiyed et al., 2025) and on the original publication of RAG (Lewis et al., 2020), this work is different from past work with metadata as filter (Poliakov & Shvai, 2024), metadata as keyword source to retrieve datasets (Hayashi et al., 2024), metadata comprehension (Shaik et al., 2024), or metadata abstracts to construct expanded search queries in systematic reviews (Li et al., 2024). There are similar elements to be explored further such as using bibliometric values for ranking (Li et al., 2024), or choice of embedding models for descriptive metadata (Hayashi et al., 2024), or even directly drawing metadata from database snapshots using generated SQL queries (Shaik et al., 2024). While there are multiple RAG architectures to address various query types, there is a lack of case studies for the LIS community that explicitly show the connection between bibliographic metadata characteristics and errors and RAG configuration decisions.

As LLM-powered applications proliferate in the academic environment and some publishers are actively removing information from public access as a strategic withdrawal from publicly available resources such as Crossref and OpenAlex[2], an understanding of how metadata as an external knowledge source for RAG applications and how metadata contents affect RAG outcomes is beneficial for the LIS community in three ways. One, it characterizes the current state of the title and abstract element and expands the concept of quality. Two, it assesses how these elements are transformed (if at all) when moving from RA to aggregator which helps the community understand limitations in their sources of data. Three, it may provide insights into how RAG components work, their limitations, and may provide critical awareness of how data source decisions impact RAG systems which is important as the academic infrastructure is rapidly deploying such systems.

I see the following opportunities:

1. investigate works from Crossref, characterizing metadata elements (DOI, title, abstract, language) and identify quality problems,
2. compare the same metadata elements in shared works between Crossref and OpenAlex, to understand what quality issues may be introduced with transformations from their original source, and
3. explore the effects of metadata characteristics or errors on configuration decisions when using a discipline specific metadata corpus as an external knowledge source in a RAG Q&A application as an exploratory case study.

Research Questions

Assuming metadata elements of the DOI, title, and abstract can be used as grounding for a RAG Q&A application, the purpose of this research is to identify the characteristics in metadata content that may affect RAG configuration decisions, how metadata sources compare for these elements, and explore what this looks like in action with a case study.

1. What is the characterization of the metadata elements from a random sample of Crossref metadata?
2. What is the type and quantity of errors, enhancements, or transformations when a shared corpus in Crossref and OpenAlex are compared?
3. How do metadata characteristics and errors affect retriever performance and generated responses?

Study design

There are three parts to answer each RQ.

[2] code change to OpenAlex that makes abstracts for both Springer and Elsevier = None: https://github.com/ourresearch/openalex-guts/commit/b85b3bc77cf9c0f3bd162426a2ba0dacdc951065

Part 1:

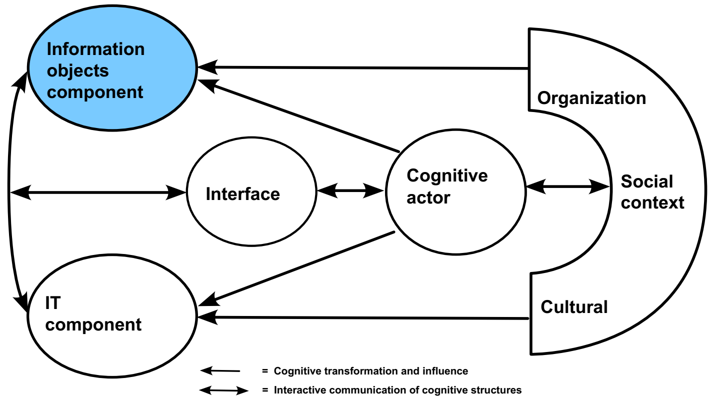

Figure 3: Adapted IS&R model for Part 1 showing the focal area for descriptive study.

Part 1 will be a descriptive study on the dimensions of the Information objects component examining the metadata as a document. A generalizable sampling of works (journal articles and conference proceedings) from Crossref will be analyzed for quality by identifying incorrect values, incorrect elements, missing information, and inconsistent value representation (Yasser, 2011, p. 51) and characterize content by identifying patterns in text content. Elements to be examined will include DOI, title, and abstract. Their quality will be assessed by incorrect values, incorrect elements (i.e., "Title" in the article title metadata), missing information, and inconsistent value representation (i.e., using cc by, instead of CC-BY or titles in ALL CAPS). Characterization of metadata element contents (such as title, and abstract) will include counts of tokens, special characters, numerals, non-text elements, and non-relevant content or characters. Metadata elements will be with respect to the permissible limits of the Crossref 5.4.0 metadata schema (Feeney, 2025), for example, face markup languages not permitted.

Following the recommendations for study design in Ingwersen & Järvelin (2005), in Part 1, controlled variables will be established through data collection from the REST API and observed variables of the document sign language and content will be documented and quantified from the metadata content.

Table 2.: Design cube dimensions of the Document and source component for Part 1

Document and source

Controlled variables	Observed variables
Document structure	Document sign language
Document types	Document content
Information type in document	
Communication function	
Temporal aspects	
Layout and style	

Figure 3 shows an overview of the process for Part 1 with analysis of the full sample set and the subset.

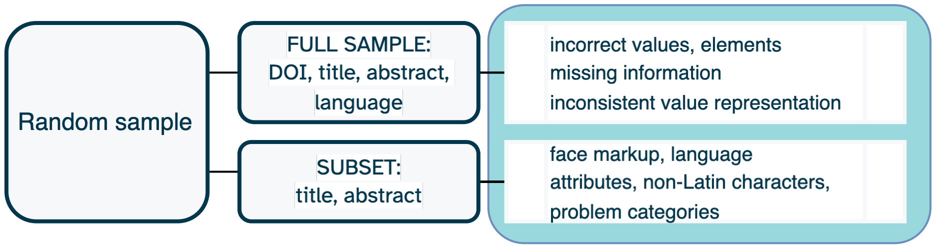

Figure 4: Part 1 process diagram for full sample (top) and subset sample (bottom).

Part 2:

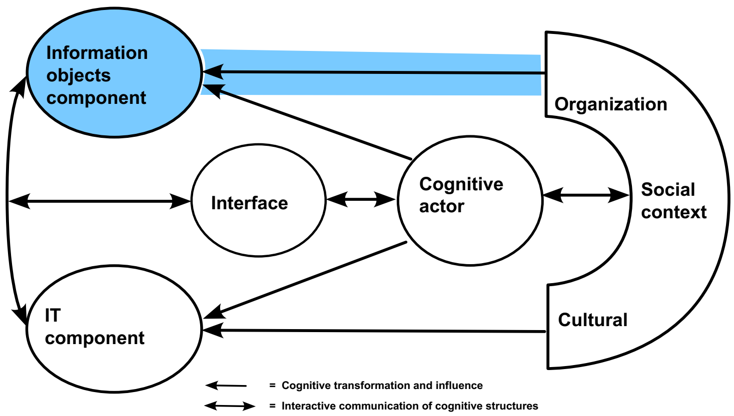

Figure 5: Adapted IS&R model for Part 2 showing focal areas of the Information object component and organizational/social/cultural influence link.

Part 2 will examine the influence of the metadata source on the content of the title and abstract. Using a shared corpus (Culbert et al., 2024) that is co-occurring in Crossref and OpenAlex, I will identify how metadata contents compare to understand what quality problems exist with transformations from their original Crossref source. The shared corpus will be matched on their DOI, acknowledging the limitations of this method (Culbert et al., 2024). A map of the Crossref schema definitions and the aggregator's schema will be created using open-source code (*Ourresearch/OpenAlex*, 2023/2024) similar to past work on document types by (Haupka et al., 2024). The same metadata elements from Part 1 will be used to identify transformations of the publisher-deposited data (from the RA) and from the aggregator. Differences will be quantified and coded into error types (Shi et al., 2025; Yasser, 2011). Patterns identified in Part 1 will also be used on the full dataset to determine differences between the two sources.

Using the design cube to identify variables, the study focuses on the dependent variables of the content and language of the title and abstract which will be measured by changes introduced by the independent variable of the metadata source. Controlled variables extend to include the organizational component to account for the influence link of the metadata source, specifically the dependencies created by schema restrictions and ingest processes.

Table 3: Design cube variables (from Ingwersen & Järvelin) for Part 2, arranged by component and dimensions.

Organizational Task Component	Document and source

Controlled variables	Controlled variables	Dependent variables	Independent variable
Task dependencies	Document structure	Document sign language	Data source
	Document types	Document content	
	Information type in document		
	Communication function		
	Temporal aspects		
	Layout and style		
	Document metadata		

Figure 5 below shows the process for Part 2, which is split into two sets of analysis once DOIs have been matched between the two datasets. The entire shared corpus is analysed for differences with the same quantification of the OpenAlex data as in Part 1. The subset of the shared corpus will also compare titles and abstracts that have been flagged as not exact to document differences.

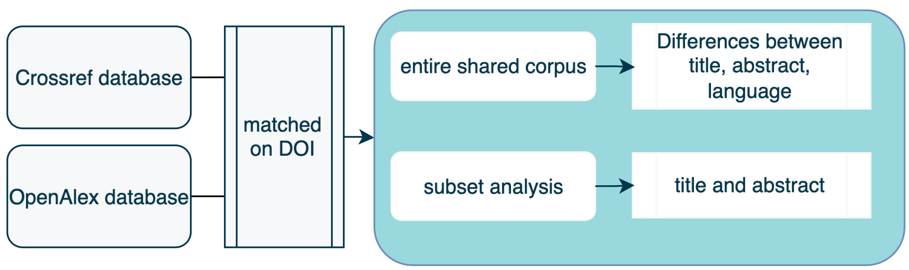

Fig 5: Part 2 process to compare metadata content from two sources.

Part 3:

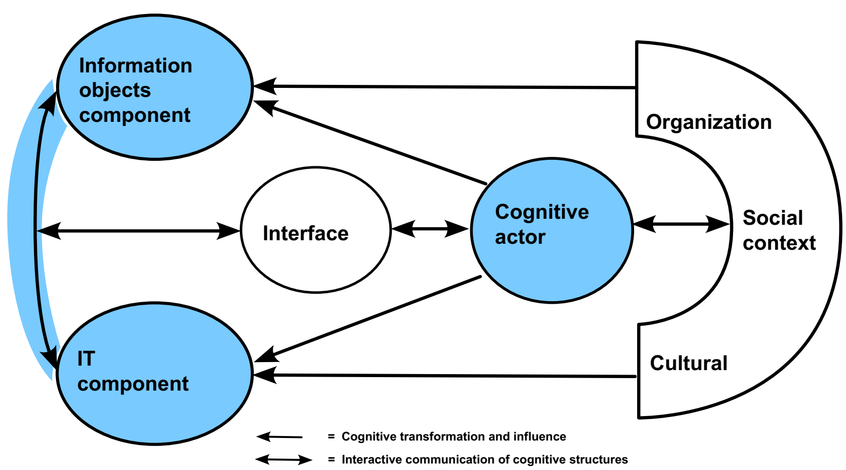

Figure 6: Adapted IS&R model for Part 3 showing the information object, IT, and actor components selected as sites of controlled, independent, or observed variables.

A case study using a constructed corpus of metadata will be used as the external knowledge source in a basic RAG application. Characteristics and errors identified in Part 1 will be used to modify a control corpus to create additional conditions, such as encodings, or multilingual text. A golden set of queries with identified true positive documents will be used on the RAG application. Analysis will examine retriever performance and generated outcomes. Typically, RAG studies qualify outputs using a benchmark dataset, such as quac[3], SciFact[4] or SciDocs[5], which may include queries, contexts, and ground truth answers. These models were built with different purposes and do not necessarily have the full abstracts or queries as part of the data. There have been other studies that have created their own dataset and made it available as public, such as Antal & Buza's dataset of thesis abstracts (Antal & Buza, 2025), Feb4RAG which was developed based on existing datasets (Wang et al., 2024), or AMAQA (Bruni et al., 2025). PubMedQA (PQA-L) contains questions and multiple contexts extracted from scientific papers along with long and short answers (Jin et al., 2019). The problem with pre-existing datasets is that they were developed for very specific research questions or to evaluate specific RAG architectures. SciFact contains claims and queries, but not whole abstracts and each question only return one text. SciDocs only includes the title and was designed to test a specific embedding model, SPECTER, not test retrieval. Antal & Buza's dataset contains thesis abstracts but only has one query for one document and introduces the confounding variable of English syntax errors. Feb4RAG integrates subsets of many BEIR datasets, including SciFact and SciDocs, but once again, only provide one query for one document making retrieval scoring impossible. PubMedQA, specifically the PQA-L expert labelled dataset of 1000 questions, comes close as the dataset could be filtered for those questions with multiple contexts, but as it was developed for evaluating the reasoning capabilities of models, the extracted texts are short, include irrelevant text, and was intended for yes/no answers to questions. This could be useful for response evaluation but not retriever evaluation. However, these previous studies provided processes for developing their own datasets including using LLM-assist for question-answer pair creation along with human curating and annotating for final evaluation. Following their processes in addition to the guidance from Teixeira de Lima et al. (2025) for dataset creation will provide the rigor necessary for creating my own dataset. They reinforce that knowing your data 'is an important step…to properly evaluate and optimize their own systems' (Teixeira de Lima et al., 2025, p. 45).

In table 3 below, the components of the IR engines/IT, Document and source, and Actor are shown with dimensions used in Part 3. Controlled variables are included for the actor due to the use of queries which assumes domain and IS&R knowledge and assumes a stage in a simulated work task. Controlled variables also are accounted for in the Document and source dimension and in the IR/IT dimension. A single independent variable, the document sets that serve as the external knowledge for the RAG, will be manipulated to observe changes in the observed variables affecting document vectorization and response generation.

[3] https://huggingface.co/datasets/allenai/quac
[4] https://huggingface.co/datasets/BeIR/scifact
[5] https://github.com/allenai/scidocs

Table 4: Components and Dimensions for Part 3 as controlled, independent, and observed variables.

Actor	**Document and source**		**IR engines and IT components**	
Controlled variable	**Controlled variable**	**Independent variable**	**Controlled variable**	**Observed variable**
Domain knowledge	Document structure	Document content	Best match models	Use of NLP for document vectorization
IS&R knowledge	Document types		Degree of document structure and context used (chunking)	Response generation
Stage in WT execution	Information type in document			
	Communication function		Use of weights in document vectorization	
	Temporal aspects		Degree of required structure and context used	
	Document sign language		Use of NLP for request vectorization	
			Use of weights in query vectorization	

Figure 7 below shows how the three datasets, (as the independent variable) will be sent through the RAG pipeline. Evaluation metrics precision, recall, accuracy, balanced accuracy, and document scores will measure the effect on the retriever which uses the vectorization of the documents and the query to determine a best match for similarity. The generated response will be measured for faithfulness and answer relevance.

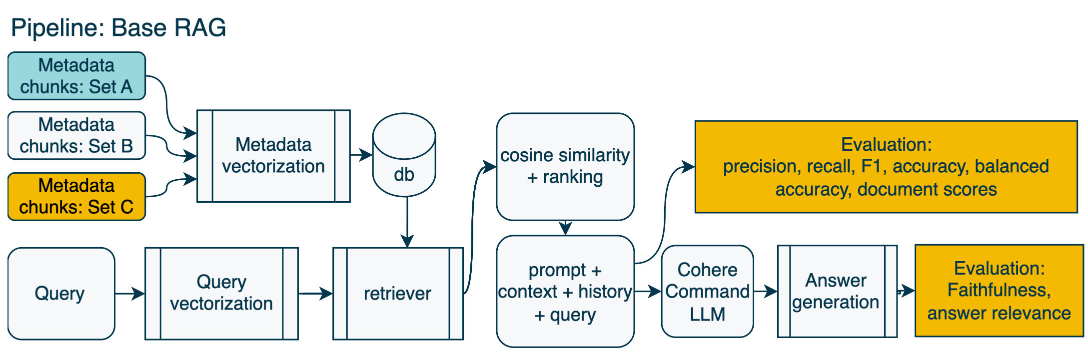

Fig 7: RAG pipeline for Part 3 showing evaluation metrics with respect to the retriever and the generation steps.

In keeping with the actor-centred intention of Ingwersen & Järvelin's IS&R model, particularly with respect to multiple influences on relevance judgements and in recognizing the limitations of existing datasets, I chose to create my own benchmark following the question taxonomy of Teixera de Lima et al., (2025) using single fact and summary question types with a simple syntax. Questions were created based on the methods by Bruni et al., (Bruni et al., 2025) with LLM-assist followed by human validation of question/answer pairs. Human annotator pairs and human-LLM annotator pairs will be validated using Cohen's kappa for agreement. For LLM-assisted work in question generation and annotation, prompts from Es et al., (2024) will guide prompt development on Cohere models.

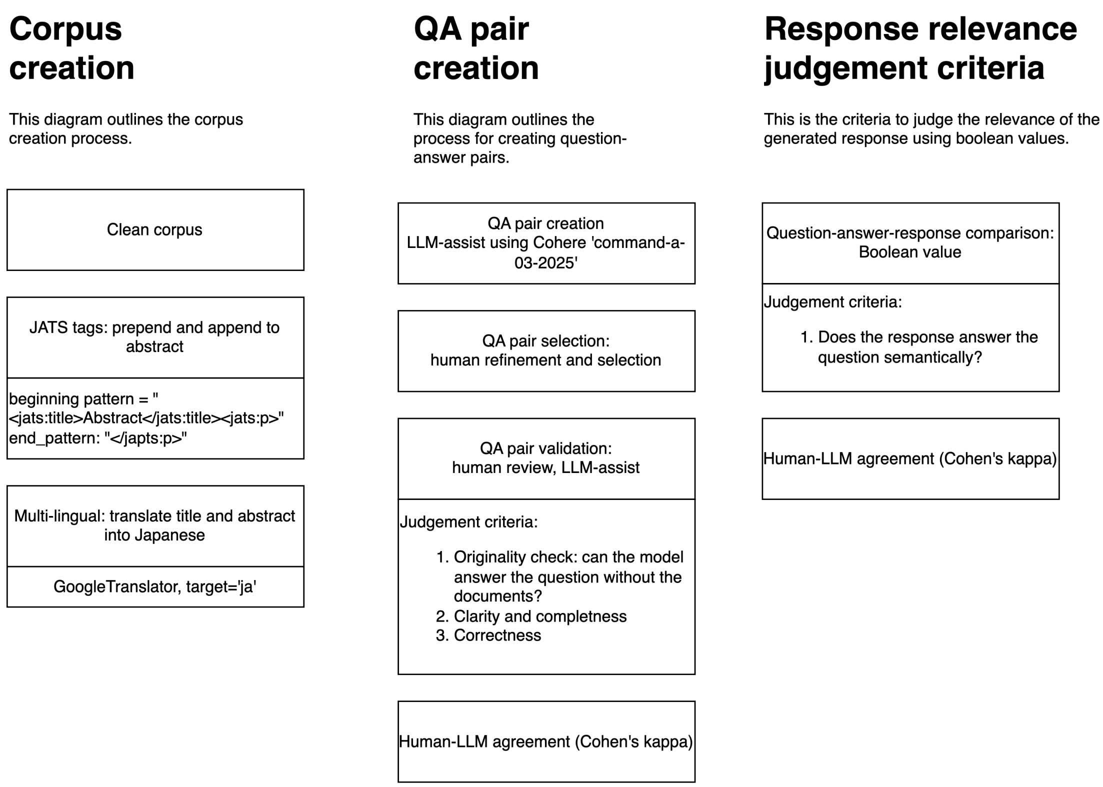

Figure 7: Processes for corpus creation, QA pair creation, and response relevance judging criteria. Where agreement will be calculated, Cohen's kappa will be used as performed in prior research (Wang et al., 2024).

Retrieval will be measured by precision, recall, F1, accuracy, and balanced accuracy scores will measure the effects of metadata characteristics or errors introduced to the dataset. Responses will be measured for faithfulness (inclusion of retrieved results in the response) and answer relevance (did the response answer the query) using both a human and LLM-as-judge approach as in prior work (Antal & Buza, 2025; Wang et al., 2024).

Potential impacts of the research

This study seeks to contribute, as novel methodological contributions, to the information science literature in the following areas. In Part 1, I identify characteristics of metadata (title and abstract) from a registration agency source, analysing the content of the text to identify patterns or errors that may impact NLP processes. While analysis of the content normally is beyond scientometrics, it does pertain to local, regional and publisher practices and may be further affected by proposed enrichment processes. In Part 2, I characterize the types and frequency of errors that may arise from transformations from source to aggregators and investigate how these characteristics in the title and abstract may affect RAG performance. Comparison of databases is normal within scientometrics and information science, yet this purpose is a novel and needed critical analysis of ingest methods for the title and abstract. In Part 3, I use existing methods but create a novel dataset for the purpose of analyzing retrieval and generation.

Empirical contributions include the quantitative results from Part 1 and 2, as well as classification of characteristics/errors into patterns that were applied to the full dataset. These may be easily used by the community as a further practical contribution. Findings from Part 3 may be of benefit to those in the RAG/IR community as empirical contributions, particularly for identification of another threat to embedding and generator model robustness.

As a theoretical contribution, I have refined the dimensions of the IS&R model to account for understanding the influences of RAG technology within the IT component. This research does not refine all dimensions of the IS&R components, yet I recognize that future work should investigate tasks within the organizational component and access and interaction components to adapt to RAG. While the IS&R model is intended to be media independent, the IT component has not aged as well and has been refined to work within the context of RAG applications.

Summary

In the context of how metadata may be used as a source of information for RAG-based applications, the purpose of this study is to investigate characteristics and errors of metadata elements in a generalizable sample, how sources of metadata compare for such characteristics/errors, and explore how these characteristics may affect RAG performance by measuring retriever performance and generated responses. Recent research showed that typos, irrelevant, contradicting, or counterfactual content can affect RAG outcomes. Given current conversations around metadata enrichment and bibliographic databases beginning to consider this a possibility, it is unknown to the community what enrichment may mean for content changes within the title and abstract particularly for a new paradigm of using metadata as an external knowledge source. Knowing how characteristics or errors affect this potential new use of metadata may affect RAG developers for decisions on cleaning or source. It may also help database management consider improvements on metadata ingest, particularly from enrichment processes.

References:

Alperin, J. P., Portenoy, J., Demes, K., Larivière, V., & Haustein, S. (2024). *An analysis of the suitability of OpenAlex for bibliometric analyses* (No. arXiv:2404.17663). arXiv. https://doi.org/10.48550/arXiv.2404.17663

Antal, M., & Buza, K. (2025). Evaluating Open-Source LLMs in RAG Systems: A Benchmark on Diploma Theses Abstracts Using Ragas. *Acta Universitatis Sapientiae, Informatica*, *17*(1), 5. https://doi.org/10.1007/s44427-025-00006-3

Belkin, N. J. (1984). Cognitive models and information transfer. *Social Science Information Studies*, *4*(2), 111–129. https://doi.org/10.1016/0143-6236(84)90070-X

Bruni, D., Avvenuti, M., Tonellotto, N., & Tesconi, M. (2025). *AMAQA: A Metadata-based QA Dataset for RAG Systems* (No. arXiv:2505.13557). arXiv. https://doi.org/10.48550/arXiv.2505.13557

Céspedes, L., Kozlowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimy, S., Ayeni, P., Filali, S., Li, B., & Larivière, V. (2024). *Evaluating the Linguistic Coverage of OpenAlex: An Assessment of Metadata Accuracy and Completeness* (No. arXiv:2409.10633). arXiv. https://doi.org/10.48550/arXiv.2409.10633

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), Article 16. https://doi.org/10.1609/aaai.v38i16.29728

Cho, S., Jeong, S., Seo, J., Hwang, T., & Park, J. C. (2024). *Typos that Broke the RAG's Back: Genetic Attack on RAG Pipeline by Simulating Documents in the Wild via Low-level Perturbations* (No. arXiv:2404.13948). arXiv. https://doi.org/10.48550/arXiv.2404.13948

Cioffi, A., Coppini, S., Massari, A., Moretti, A., Peroni, S., Santini, C., & Shahidzadeh Asadi, N. (2022). Identifying and correcting invalid citations due to DOI errors in Crossref data. *Scientometrics*, *127*(6), 3593–3612. https://doi.org/10.1007/s11192-022-04367-w

Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). *Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus* (No. arXiv:2401.16359). arXiv. https://doi.org/10.48550/arXiv.2401.16359

datacite. (2025, November 10). *Request for comment: Exploring DataCite metadata enrichments · datacite/datacite-suggestions · Discussion #209*. GitHub. https://github.com/datacite/datacite-suggestions/discussions/209

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, *5*(1), 31–49. https://doi.org/10.1162/qss_a_00286

Eck, N. J. van, & Waltman, L. (2022). *Crossref as a source of open bibliographic metadata*. OSF. https://doi.org/10.31222/osf.io/smxe5

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. In N. Aletras & O. De Clercq (Eds.), *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://doi.org/10.18653/v1/2024.eacl-demo.16

Färber, M., Braun, C., Popovic, N., Saier, T., & Noullet, K. (2022, April 10). Which Publications' Metadata Are in Which Bibliographic Databases? A System for Exploration. *BIR 2022: 12th International Workshop on Bibliometric-Enhanced Information Retrieval*. ECIR2022, Hybrid.

Feeney, P. (2025, March 19). *Version 5.4.0 metadata schema update now available* [Blog]. Crossref. https://doi.org/10.13003/325070

Han, B., Susnjak, T., & Mathrani, A. (2024). Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview. *Applied Sciences*, *14*(19), 9103. https://doi.org/10.3390/app14199103

Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., & Mayr, P. (2024). *Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, Pubmed and Semantic Scholar* (No. arXiv:2406.15154). arXiv. https://doi.org/10.48550/arXiv.2406.15154

Hayashi, T., Sakaji, H., Dai, J., & Goebel, R. (2024). *Metadata-based Data Exploration with Retrieval-Augmented Generation for Large Language Models* (No. arXiv:2410.04231). arXiv. https://doi.org/10.48550/arXiv.2410.04231

Hendricks, G. (2025, November 7). *Crossref's Perspective on Open Metadata Enrichment*. COMET. https://www.cometadata.org/blog/crossrefs-perspective-on-open-metadata-enrichment

Ingwersen, P. (1992). *Information retrieval interaction*. Taylor Graham. https://peteringwersen.info/publications/0060_ingwersen_iri.pdf

Ingwersen, P. (1996). COGNITIVE PERSPECTIVES OF INFORMATION RETRIEVAL INTERACTION: ELEMENTS OF A COGNITIVE IR THEORY. *Journal of Documentation*, *52*(1), 3–50. https://doi.org/10.1108/eb026960

Ingwersen, P., & Järvelin, K. (2005). *The Turn Integration of Information Seeking and Retrieval in Context* (1st ed. 2005). Springer Netherlands : Imprint : Springer.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). *PubMedQA: A Dataset for Biomedical Research Question Answering* (No. arXiv:1909.06146). arXiv. https://doi.org/10.48550/arXiv.1909.06146

Kang, S.-H., & Kim, S.-J. (2023). M-RAG: Enhancing Open-domain Question Answering with Metadata Retrieval-Augmented Generation. 한국정보통신학회논문지, *27*(12), 1489–1500. https://doi.org/10.6109/jkiice.2023.27.12.1489

Kramer, B., & de Jonge, H. (2022). The availability and completeness of open funder metadata: Case study for publications funded by the Dutch Research Council. *Quantitative Science Studies*, *3*(3), 583–599. https://doi.org/10.1162/qss_a_00210

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. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Li, Y., Zhao, J., Li, M., Dang, Y., Yu, E., Li, J., Sun, Z., Hussein, U., Wen, J., Abdelhameed, A. M., Mai, J., Li, S., Yu, Y., Hu, X., Yang, D., Feng, J., Li, Z., He, J., Tao, W., … Tao, C. (2024). RefAI: A GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. *Journal of the American Medical Informatics Association*, *31*(9), 2030–2039. https://doi.org/10.1093/jamia/ocae129

Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., & Singh, S. (2022). *Entity-Based Knowledge Conflicts in Question Answering* (No. arXiv:2109.05052). arXiv. https://doi.org/10.48550/arXiv.2109.05052

Marjanović, S. V., Yu, H., Atanasova, P., Maistro, M., Lioma, C., & Augenstein, I. (2024). *DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models* (No. arXiv:2407.17023). arXiv. https://doi.org/10.48550/arXiv.2407.17023

Mongeon, P., Hare, M., Krause, G., Marjoram, R., Riddle, P., Toupin, R., & Wilson, S. (2025). Investigating Document Type Discrepancies between OpenAlex and the Web of Science. *Proceedings of the Annual Conference of CAIS / Actes Du Congrès Annuel de l'ACSI*. https://doi.org/10.29173/cais1943

Mongeon, P., Hare, M., Riddle, P., Wilson, S., Krause, G., Marjoram, R., & Toupin, R. (2025). *Investigating Document Type, Language, Publication Year, and Author Count Discrepancies Between OpenAlex and Web of Science* (No. arXiv:2508.18620). arXiv. https://doi.org/10.48550/arXiv.2508.18620

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Mugabushaka, A.-M., van Eck, N. J., & Waltman, L. (2022). Funding COVID-19 research: Insights from an exploratory analysis using open data infrastructures. *Quantitative Science Studies*, *3*(3), 560–582. https://doi.org/10.1162/qss_a_00212

Ourresearch/OpenAlex. (2024). [Computer software]. OurResearch. https://github.com/ourresearch/OpenAlex (Original work published 2023)

Pawlick-Potts, D. (2022). *Is anybody in there?: Towards a model of affect and trust in human – AI information interactions*. https://doi.org/10.18452/25258

Poliakov, M., & Shvai, N. (2024). *Multi-Meta-RAG: Improving RAG for Multi-Hop Queries using Database Filtering with LLM-Extracted Metadata* (No. arXiv:2406.13213). arXiv. https://doi.org/10.48550/arXiv.2406.13213

Saiyed, Z., Orian, C., Riddle, P., Prashanth Kanagaraj, G., Krause, G., Toupin, R., & Brooks, S. (2025). RAG Based Navigation of the World Ocean Assessment II. *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, 282–290. https://aclanthology.org/2025.konvens-2.22.pdf

Santos, E. A. dos, Peroni, S., & Mucheroni, M. L. (2023). An analysis of citing and referencing habits across all scholarly disciplines: Approaches and trends in bibliographic referencing and citing practices. *Journal of Documentation*, *79*(7), 196–224. https://doi.org/10.1108/JD-10-2022-0234

Saracevic, T. (1996). Modeling Interaction in Information Retrieval (IR): A Review and Proposal. *Proceedings of the ASIS Annual Meeting*, *33*, 3–9.

Savolainen, R. (2018). Pioneering models for information interaction in the context of information seeking and retrieval. *Journal of Documentation*, *74*(5), 966–986. https://doi.org/10.1108/JD-11-2017-0154

Schares, E. (2024). Comparing Funder Metadata in OpenAlex and Dimensions. *OpenISU*. https://doi.org/10.31274/b8136f97.ccc3dae4

Scheidsteger, T., & Haunschild, R. (2023). Which of the metadata with relevance for bibliometrics are the same and which are different when switching from Microsoft Academic Graph to OpenAlex? *Profesional de La Información*, *32*(2), Article 2. https://doi.org/10.3145/epi.2023.mar.09

Schlosser, M. (2016). Write up! A Study of Copyright Information on Library-Published Journals. *Journal of Librarianship and Scholarly Communication*, *4*(0), Article 0. https://doi.org/10.7710/2162-3309.2110

Shaik, K., Wang, D., Zheng, W., Cao, Q., Fan, H., Schwartz, P., & Feng, Y. (2024). S3LLM: Large-Scale Scientific Software Understanding with LLMs Using Source, Metadata, and Document. In L. Franco, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Computational Science – ICCS 2024* (pp. 222–230). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-63759-9_27

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Schärli, N., & Zhou, D. (2023). Large Language Models Can Be Easily Distracted by Irrelevant Context. *Proceedings of the 40th International Conference on Machine Learning*, 31210–31227. https://proceedings.mlr.press/v202/shi23a.html

Shi, J., Nason, M., Tullney, M., & Alperin, J. P. (2025). Identifying Metadata Quality Issues Across Cultures. *College & Research Libraries*, *86*(1), Article 1. https://doi.org/10.5860/crl.86.1.101

Teixeira de Lima, R., Gupta, S., Berrospi Ramis, C., Mishra, L., Dolfi, M., Staar, P., & Vagenas, P. (2025). Know Your RAG: Dataset Taxonomy and Generation Strategies for Evaluating RAG Systems. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert, K. Darwish, & A. Agarwal (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track* (pp. 39–57). Association for Computational Linguistics. https://aclanthology.org/2025.coling-industry.4/

Wan, A., Wallace, E., & Klein, D. (2024). *What Evidence Do Language Models Find Convincing?* (No. arXiv:2402.11782). arXiv. https://doi.org/10.48550/arXiv.2402.11782

Wang, S., Khramtsova, E., Zhuang, S., & Zuccon, G. (2024). FeB4RAG: Evaluating Federated Search in the Context of Retrieval Augmented Generation. *Proceedings of the 47th International ACM SIGIR*

Conference on Research and Development in Information Retrieval, 763–773. https://doi.org/10.1145/3626772.3657853

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation*, *55*(3), 249–270. https://doi.org/10.1108/EUM0000000007145

Yasser, C. M. (2011). An Analysis of Problems in Metadata Records. *Journal of Library Metadata*, *11*(2), 51–62. https://doi.org/10.1080/19386389.2011.570654

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics*, *129*(10), 5869–5891. https://doi.org/10.1007/s11192-023-04923-y