Generative AI-augmented and User-centric Research Data Discovery and Reuse

by

Lizhou Fan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in The University of Michigan
2024

Doctoral Committee:

Associate Professor Libby Hemphill, Chair
Professor Anne Gilliland, UCLA
Professor H. V. Jagadish
Professor Margaret Levenstein

Lizhou Fan
lizhouf@umich.edu
ORCID iD: 0000-0002-7962-9113

Dedication

To my beloved wife, Huizi Yu, for her unwavering love, support, and patience throughout this journey, and to my long-term research partner, Huizi Yu, whose collaboration since our first paper has been instrumental in shaping my research and achievements. This dissertation stands as a testament to your enduring contributions to my life and work.

Acknowledgements

I am profoundly grateful to my PhD advisor in UMSI, Libby Hemphill, for her invaluable trust and guidance, which have been fundamental to my academic and personal development throughout this journey. I extend my heartfelt thanks to my undergraduate research supervisors in UCLA, Anne Gilliland and Todd Presner, whose expertise in archival science and digital humanities laid the groundwork for my research pursuits. Additionally, I am indebted to Xin Ma and Yongfeng Zhang for their mentorship in biomedical informatics, computer vision, and natural language processing, which has significantly shaped the next step of my research direction.

My sincere gratitude also goes to my research collaborators—Sara Lafia, Lingyao Li, Wenyue Hua, Jiayan Zhou, and Xiang Li, among many others. Your insights and unwavering dedication have profoundly enriched my research. I am equally thankful to the staff at the Inter-university Consortium for Political and Social Research (ICPSR)—David Bleckley, Elizabeth Moss, Bing Zhang, and Aalap Doshi. Collaborating with each of you has been a genuinely enriching experience that I will treasure.

Lastly, I must express my deepest appreciation to my family, especially my parents, Li Zhu and Kai Fan, for their constant support and encouragement. Your faith in my abilities and your unconditional love have been my steadfast anchors, motivating me through every challenge.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

This dissertation addresses the challenge of enhancing research data discovery and reuse in the face of escalating data volume and complexity. Traditional metadata-driven search tools often fall short in providing nuanced context and interdisciplinary connections critical for efficient scientific exploration and collaboration. To address these limitations, we developed the Generative AI-augmented and User-centric Data Search (GAUDS) system, which integrates Large Language Models (LLMs) and Scholarly Knowledge Graphs (SKGs) to parse natural language queries and visualize data relationships, thereby fostering a deeper understanding of available research resources.

The study details the development and implementation of the GAUDS system, including the conceptualization of the guiding principles Connectivity, Effectiveness, Visibility and Interactivity (CEVI) that support and evaluate the discovery and reuse of research data. It further explores the construction of the ICPSR Health and Medical Scholarly Knowledge Graph (IHSKG), which represents complex connections in research data and prototypes interdisciplinary reuse potentials. The abilities of LLMs to perform complex reasoning were assessed, informing the system's ability to understand and manipulate large datasets effectively.

The development of the GAUDS system, informed by insights gained from prototyping and evaluating user-centric utility, leads to a comprehensive analysis of focus group feedback. This feedback evaluates the system’s impact on enhancing data discoverability and usability. The GAUDS system, by providing effective navigation aids, relevant dataset suggestions, and contextualized reuse guides, not only enhances user engagement and satisfaction, but also demonstrates the transformative potential of generative AI in specialized academic domains such as health and medical research. This research contributes to the fields of information retrieval and data management by proposing a novel approach that combines human-curated knowledge graphs with generative AI algorithms to significantly improve data discovery and reuse. Future work will aim to productionize the GAUDS system, expand its scalability across different domains, and explore its broader potential to support open science initiatives.

Chapter 1

Introduction

Research data discovery and reuse are fundamental aspects of scientific research, playing a crucial role in driving innovation [1], fostering collaboration [2], and ensuring the reproducibility of scientific findings [3], [4]. As the volume and complexity of academic literature and research data continue to grow at an unprecedented rate [5], [6], the need for effective data search and recommendation strategies becomes increasingly important. Meanwhile, open science and open data guidelines emphasize the importance of discoverability and reusability of research objects [7], [8]. These guidelines aim to promote the sharing of scientific findings and the repurposing of research data for secondary use, thereby maximizing the value and impact of research investments.

To ensure data discovery and reuse, one effective approach is to leverage the efforts of data management units that offer data search and recommendation services. Institutional repositories, in particular, serve a critical data management role within the academic and research landscape, as they provide a centralized and accessible platform for storing, sharing, and discovering research data [9], [10]. By utilizing these services, researchers and other data users expect to locate relevant data, avoid duplicating efforts, and accelerate the pace of scientific progress.

Some institutional repositories and data archives maintain high data management standards with professional curators and bibliographers [11], [12], ensuring that research data is accurately described, preserved, and made accessible to researchers. However, the curation of research data is a labor intensive and time consuming process [13]. Many other archives and repositories tend not to have enough support and need computational approaches to augment and accelerate metadata production and indexing services for data discovery. Moreover, due to the uneven support from different data funding agencies and owners [14]–[16], the distribution of data curation time and resources can be different for each dataset. Thus, the discoverability and reusability of different datasets, even hosted by the same institutional repository, can be different.

While data providers such as institutional repositories have made significant strides in offering standardized metadata and search tools to data users, there remains a gap in providing users with the nuanced context and domain expertise required for effective dataset discovery and reuse [17], [18]. Data management standards, such as the FAIR principles [19], [20], ensure that data is machine-readable and can be easily shared and reused. However, these standards often fall short of offering users a holistic understanding of the data, including its provenance, quality, and relevance to specific research questions [21]. This gap in context and expertise can make it difficult for researchers to effectively discover and reuse research data [22], [23]. Analysis of researchers' data reuse behaviors shows that they may ”reverse engineer” or ”back into data analysis” by starting from data-related literature to understand data [24], requiring a high level of expertise and domain knowledge. As a result, they may struggle to determine whether a particular dataset is relevant to their research question, or whether the data is of sufficient quality for their intended use. This can lead to missed opportunities for data reuse, or worse, the use of inappropriate or low-quality data in scientific research.

In addition to the context and expertise challenges, on the data management end, there are also current issues that negatively impact data discovery and reuse. First, recognizing data ownership is essential as data might be easily findable, yet stakeholders remain unidentified, posing ethical challenges [19], [25]. Second, the efficiency of data management systems can significantly impact the cost, time, and resources involved [26]. Third, a notable difficulty lies in the ease-security trade-off, where although data access is granted, low interoperability results in users losing context, thereby echoing the context challenge [27]–[32]. Furthermore, while access to data is available, its low reusability due to knowledge complexity leads to insufficient expertise among users, reinforcing the expertise challenge [29], [33]–[37]. These interconnected issues highlight the broader implications of data management that need addressing to harness the full potential of data in research and development.

To address the above challenges in data discovery and reuse, we need to develop data search and recommendation systems that go beyond traditional metadata-based matching approaches, fostering enhanced user-data interaction. Such systems aim to help users seamlessly discover data, understand the data and related metadata, and make informed decisions about reuse. To achieve this enhanced user-data interaction, Large Language Models (LLMs) and Scholarly Knowledge Graphs (SKGs) play pivotal roles. Recent advances in computing, particularly in the areas of distributed cloud computing and accelerated hardware, have enhances the capabilities of both SKGs and LLMs. Distributed cloud environments facilitate the expansion of SKGs by allowing for more extensive data integration and real-time data-user interaction [38], [39]. Meanwhile, improvements in Graphics Processing Unit (GPU) technologies have accelerated the training and deployment phases of LLMs [40], [41], making it feasible to handle larger datasets and perform more sophisticated analyses at unprecedented speeds. This synergy between advanced computing infrastructure and innovative data techniques is crucial for developing more intelligent, adaptive, and user-centric data discovery tools.

LLMs are Generative Artificial Intelligence (Generative AI) models specialized in Natural Language Processing (NLP) [42], [43]. LLMs can facilitate two-way understandability, i.e., simultaneously learning complex patterns in data and generating human-like responses [44], [45], making them ideal for supporting data search and engaging data users with high-quality generated reuse guidance. While directly applying LLMs in research data user systems is an novel approach, these models have shown proficiency in chatbot applications, adeptly handling user inquiries with responses that are both relevant and contextually appropriate [46]–[48]. In particular, LLMs can translate natural language queries into machine readable database queries, understand semantic meaning in search terms and find similar items, and contextualize generated results based on specific user input.

SKGs employ a graph-based non-relational database structure to represent scholarly knowledge, capturing intricate relationships between various entities such as datasets, publications, and researchers [49]. This structural advancement in graph-based paradigm emphasizes connections, which outperforms traditional

SQL-like database paradigms regarding nuanced exploration utility in linked data [50], [51], enabling a comprehensive scoping of context of research data by indirect relations between datasets and their metadata. By integrating LLMs with SKGs, it becomes possible to construct a system that interacts with users through natural language, instead of segmented keywords or phrases. This system also leverages the rich relational data from SKGs to provide personalized and contextually relevant data recommendations.

In this study, we develop a Generative AI-augmented and user-centric data search (GAUDS) system, powered by LLMs and SKGs and with a visual and interactive frontend. The system can maximize the usability of the information encoded in SKGs by using LLMs to convert users' natural language search queries into machine-readable information, including word embedding vectors and graph database queries (in Cypher). This translation enables the system to provide users with more accurate and relevant search results and reuse guidance. Thus, the GAUDS system offers a more intuitive and effective method for data discovery and reuse, aligning with the needs of modern research practices and adhering to open data guidelines.

The development of a generative AI-augmented and user-centric approach in research data discovery and reuse has broader implications for the field of information retrieval and data management. We expect our work to impact not just the data user of a single institutional repository, but to extend beyond a specific domain to further study general patterns in search-related knowledge discovery and reuse. By exploring innovative approaches to building SKGs, aiding users in information seeking, and employing LLMs in complex search scenarios, this research contributes to the advancement of knowledge discovery methods, facilitating more effective and efficient practices in research data management and consumption, benefiting the full data life cycle from producing to using research data.

1.1 Background and Motivations

The Inter-university Consortium for Political and Social Research (ICPSR), a international institutional repository, plays a crucial role in illuminating the complexities of data discovery and reuse in academic research [52]–[54]. Despite ICPSR's extensive repository of datasets and associated publications, the intricate connections between these resources often remain obscured to researchers due to the limitations of traditional search interfaces [11]. Researchers typically engage with ICPSR's robust web search system, built on a Solr index [55], or through third-party aggregators such as Google Dataset Search [49]. While these tools allow for efficient navigation of study-level metadata, codebooks, variables, and publications, they predominantly facilitate a metadata-centric approach to search, which can overlook the deeper, contextual relationships between datasets and scholarly outputs.

ICPSR's primary interaction portal, the 'Find Data' webpage, features user-friendly search and navigation tools such as a search box, a word cloud of popular search topics, and a list of the most downloaded datasets (see Figure 1.1). These functionalities support various user search behaviors, including direct, orienting, and scenic searches [56]. However, despite facilitating access to datasets and publications through metadata matching, the fundamental relational links among these objects are not immediately evident, necessitating additional navigational actions from users, such as following hyperlinks or toggling between interface tabs.

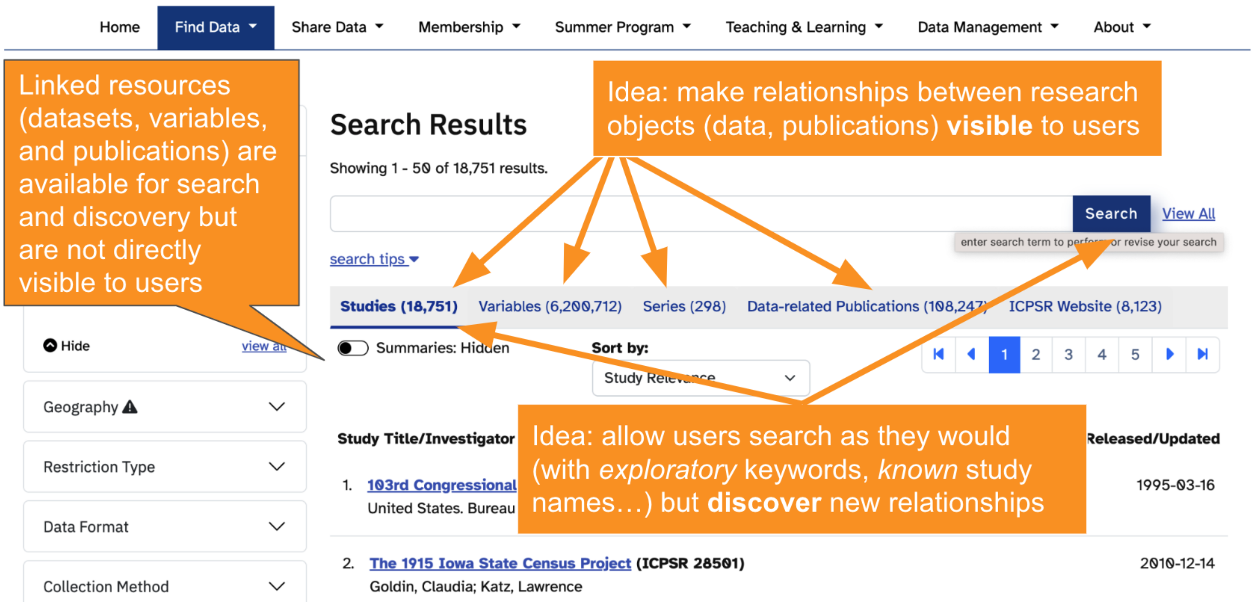

Figure 1.1: ICPSR's Find Data Webpage

The motivation behind enhancing the data discovery and reuse capabilities of ICPSR with SKGs and LLMs stems from a fundamental need to bridge the gap between simple metadata searches and a more holistic, context-driven understanding of research data. Research highlights the social-technical complexities of data search [57] and underscores the disciplinary differences that can affect data search strategies [58]. Furthermore, enhancing the search process can also positively impact researchers' intentions to access data, as indicated by studies on search effort [59]. By integrating these advanced technologies, the GAUDS system aims to simplify the data discovery process and to provide more effective and intuitive user interactions, thus significantly enhancing the efficiency and effectiveness of scholarly research.

Unlike the existing Solr-based system [55], graph databases—and by extension, SKGs—offer a transformative approach to data organization and representation. SKGs prioritize the visualization and mapping of relationships between entity types such as datasets, publications, and authors [50], [51], supporting data search systems in assisting users to understand the relations between datasets and their context, thereby enhancing data discoverability and facilitating informed data resue decisions. There is also a recognized need for data search engines to assist researchers in identifying relevant data repositories outside their primary domains [60], which are unfamiliar context to data users. Additionally, analyses from ICPSR's user search data suggest that nearly half of all queries are topically driven (Figure 1.2) [56], [61], while keyword-only search strategies often fall short in representing research topics. Using a keyword or a combination of keywords in topic search are often limited due to keywords' convoluted knowledge hierarchy and ambiguity of knowledge representation [62], [63]. These limitations of keywords reduce data search engines' capabilities to capture the non-static and complex knowledge in real-world research trends and terminologies. As such, SKG-based search approach can contribute to both backend data handling and frontend user experience by making inherent data relationships more accessible and easier to navigate.

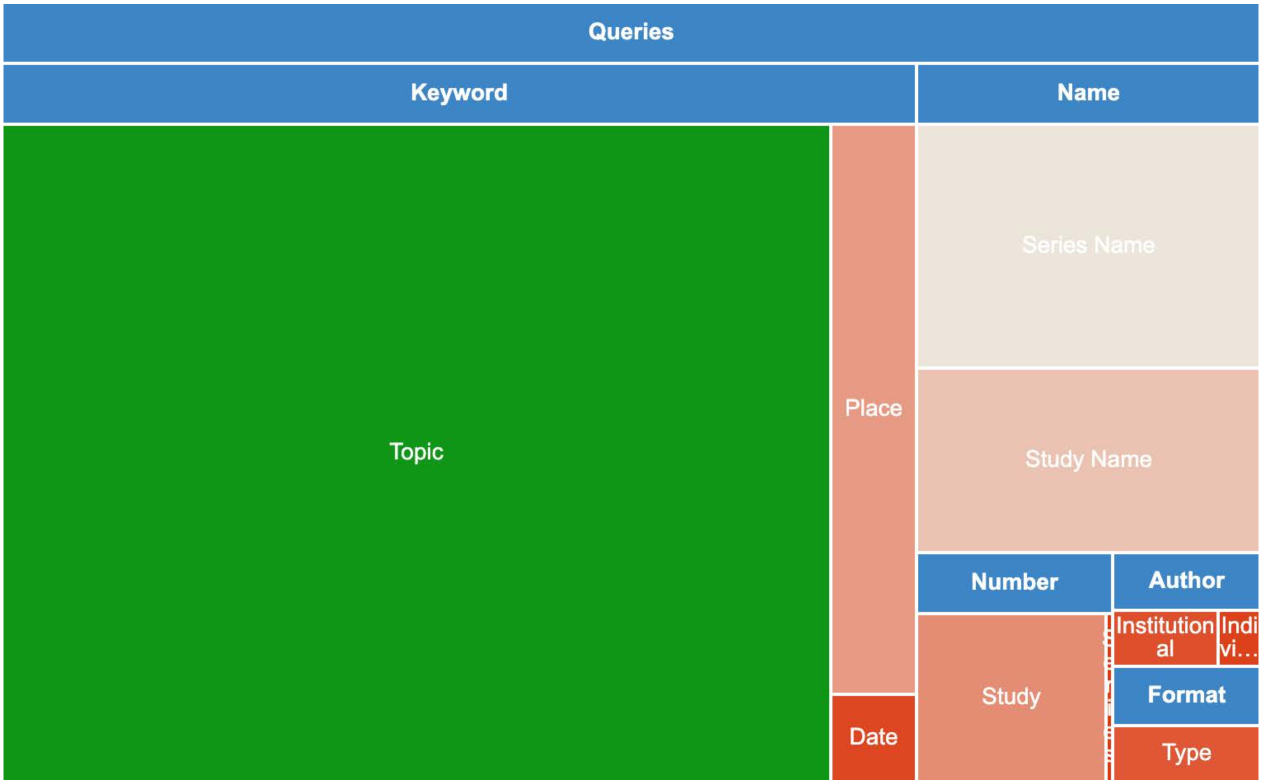

Figure 1.2: Treemap of labeled queries shows that search by topic and name were most common (figure reused with permission from [56])

The introduction of LLMs for enhancing data search and recommendation further underscores the need for advanced tools in data discovery. Traditional keyword-based searches [64]–[66], while useful, often fail to capture the nuanced demands of modern research queries. The implementation of LLMs, equipped with capabilities such as topic classification, natural language query translation, and vector-based search technologies [67]–[69], promises a more nuanced and context-aware search experience. These technologies, by converting text into semantic-rich vector representations, enable the retrieval of conceptually relevant data beyond mere keyword matches, thus addressing the limitations of current search methodologies [63]. For example, vector search technologies represent a significant shift from traditional keyword searches by leveraging the semantic relationships between words and phrases to provide richer, context-aware search results. This advancement is crucial in multidisciplinary research fields, where terminologies might differ but underlying concepts remain closely related.

To summarize, despite ICPSR's robust dataset and publication repository, the current search mechanisms, primarily based on metadata, fail to expose the deeper relational context necessary for effective data discovery and reuse. By shifting towards a more dynamic and context-aware system using SKGs and LLMs, the GAUDS system is set to revolutionize how users interact with data. By making searches simpler and more intuitive, these technologies enhance the backend data handling capabilities and enrich the frontend user experience. This is achieved by deploying LLM-based NLP algorithms that can interpret datasets' metadata and contextualize data reuse guide based on user input, thus allowing for more accurate, relevant, and dynamically generated data discovery and reuse results. Moreover, these systems are designed to seamlessly handle the complexities and disciplinary variations inherent in modern data search scenarios, improving the efficacy and efficiency of scholarly research. This integration ensures that data users are provided with intelligent recommendations and deeper insights, making navigation through extensive datasets and scholarly works more manageable and tailored to specific user needs.

1.2 Research Questions

This dissertation aims to explore advancements in research data discovery and reuse, particularly through the development of the Generative AI-augmented and User-centric Data Search (GAUDS) system. The research questions are structured to guide the design, the usability, and the initial evaluation of this system:

RQ1: How can we structure Scholarly Knowledge Graphs (SKGs) for data discovery? The goal of this research question is to enhance the discovery of research data in institutional repositories. Using the data and metadata from the Inter-university Consortium for Political and Social Research (ICPSR) as an example, we focus on the development of a comprehensive SKG with critical nodes (e.g., datasets, publications) and connections (e.g., citations, metadata links) that effectively represent the intricate relationships between data and publications. The aim is to configure these elements in a way that simplifies the data discovery process.

RQ2: How can we leverage and customize Large Language Models (LLMs) for enhanced data recommendation? This question examines how LLMs can be specifically tailored to refine data search processes and improve the quality of data recommendations provided to data users. Employing techniques such as in-context learning and model finetuning, the study aims to utilize powerful LLMs and customize them to better navigate the complexities of SKGs in a transparent and responsible way. Through LLM-based topic classification, natural language query translation, and word embedding vector search, our system is expected to lead to more precise and contextually relevant queries, significantly increasing the relevance of search results, and, by extension, enhancing the usability of LLMs in scholarly data search.

RQ3: How can we effectively communicate data recommendations and reuse suggestions to users in both written and visual formats? The third research question aims to improve the mechanisms through which data recommendations and reuse suggestions are conveyed to users, emphasizing the development of a retrieval heuristic for data use records and guidance generation that enhances both discoverability and reusability. The objective is to create end-user features in the user interface that not only present data effectively but also facilitate easy interaction with datasets, allowing users to quickly and efficiently find, comprehend and apply data in their research. This aspect of the study involves the design and implementation of a user interface that integrates principles of *Connectivity*, *Effectiveness*, *Visibility*, and *Interactivity* (CEVI). The interface leverages visual and textual cues to guide users through data selection processes, ultimately fostering more effective and informed data reuse.

1.3 Chapters Overview

To further outline the structure of the dissertation, we provide the main topics for the rest of the chapters as follows. Figure 1.3 shows an overview of the structure in this research. In this section, we outline the main structure in the rest of the dissertation and briefly describe their methods and contributions to this study[1].

Chapter 2: Literature Review. In this chapter, we first identify research gaps in the current landscape of research data management, from both curation and user perspectives. The gaps include *Data use ethics*, *Implementation efficiency*, *Ease-security trade-off*, and *Knowledge complexity*.

Chapter 3: Conceptualizing Guiding Principles and Prototyping for Data Search Systems. According to the research gaps identified in Chapter 2, we develop novel guiding principles, CEVI, for creating the GAUDS system. We then develop the SimSearch and DataChat prototypes to study the implementation of the CEVI principles and to obtain practical experience in designing data search systems.

Chapter 4: Constructing a Scholarly Knowledge Graph for Linked Research Data. Based on the theoretical developments and practical challenges discussed in Chapter 3, we strategically develop an SKG database, *ICPSR Health and Medical Scholarly Knowledge Graph* (IHSKG), for representing the complex connections in research data and prototyping interdisciplinary reuse potentials. The construction of IHSKG provides both theoretical and practical implications for RQ1.

Chapter 5: Assessing Large Language Models' Abilities for Complex Reasoning. To improve data search ease and understand complex knowledge in IHSKG, we assess the reasoning ability, and the corresponding learning ability, of LLMs and choose a set of candidates the implementation of the GAUDS system. The resulting insights of LLMs' ability in complex scenarios set the foundation for RQ2.

Chapter 6: Building the Generative AI-augmented and User-centric Data Search (GAUDS) System. Motivated by the CEVI guiding principles, learnt from the experiences in building prototypes, based on the data in IHSKG, and supported by LLMs, we create the GAUDS system in this chapter. The backend construction leverages prompt engineering and model finetuning to improve and contextualize LLMs, furthering the discussions of RQ2 and highlighting the user-centric utility in RQ3.

Chapter 7: Analyzing Focus Group Findings of the GAUDS System. In this chapter, we analyze the focus group study of the GAUDS system regarding five functionalities, including *Explore data*, *Ask data*, *Find data*, *Reuse data*, and *Show table & graph*, offering first-hand feedback and insights to address RQ3.

Chapter 8: Conclusion and Future Directions. In this final chapter, we conclude the contribution of this study. We also discuss the implications, limitations, and future directions of our research.

[1]We further show the mapping between the chapters and previous scholarly work in Appendix D

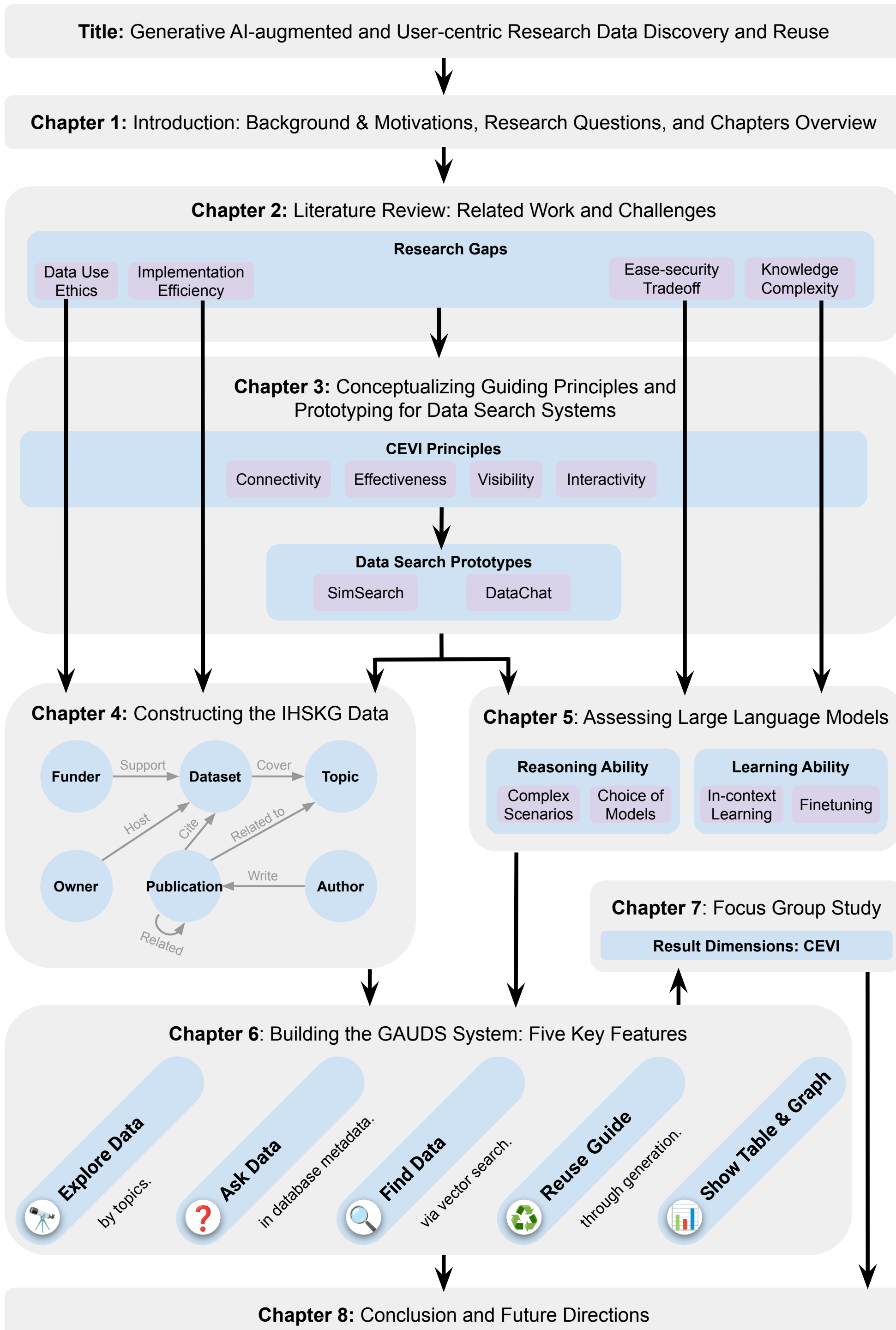

Figure 1.3: Research Overview

Chapter 2

Literature Review

Recent Research Data Management (RDM) guidelines emphasize the importance of discoverability and reusability of research objects, particularly research data [7], [8]. Data citations, which act as a link between data providers and consumers [70], provide insights into the current state of data discovery and reuse. They also underscore the collaborative efforts of providers and consumers in research data management [71]. Assessing the current status of data citations is crucial for understanding advancements in research data management and analyzing the shift in research data use and management paradigms.

The FAIR principles (Findability, Accessibility, Interoperability, and Reusability) are a set of data management guidelines that apply to RDM [20]. By following the FAIR principles, data creators and distributors can make their data more discoverable and extend their audiences. Findable and accessible data lead to efficient and effective reuse, fostering better collaboration among researchers. Additionally, complying with these principles can increase the visibility and impact of research, benefiting both individual researchers and research institutes. However, FAIR principles mainly target enhancing machine readability [20], which is not sufficient to guide production-level research data management needs. Recent extensions and critiques of the FAIR principles aim to address their shortcomings and limitations. For example, the CARE principles for data use purposes and sovereignty [72], as well as general archival principles for accessibility and usability [25], [73].

While the FAIR principles can provide informative theoretical guidelines, implementing RDM systems remains challenging due to the complexity of data discoverability and reusability and the variety of stakeholders. One approach that has gained traction in recent years is the use of Scholarly Knowledge Graphs (SKGs) for RDM. SKGs are semantic directed labeled networks of entities (nodes) linked by relations (edges) in academic research, aiming to organize scholarly knowledge and promote interoperability [50], [51]. SKGs enable the extraction of structured information from a variety of unstructured data sources, such as datasets, researchers, publications, affiliations, and funders. Users can articulate complex queries that can then be answered effectively, because of the structured nature of the ontologies in SKGs. This can reduce the amount of data preprocessing necessary when integrating data from diverse sources, resulting in more efficient data access. Moreover, the extracted information is not limited to a closed set of ontologies but can be defined by the SKGs' users, who may choose and combine needs in interdisciplinary research contexts, where data from different fields can be used out of their original scope of research. SKGs can also facilitate flexible access and navigation of interconnected research data and their metadata from diverse domains. With efficient and flexible data querying ability, SKGs help stakeholders evaluate research impacts, identify gaps in knowledge, uncover potential collaborators, and gain insights into emerging research trends [51], [74], which shed light on important RDM tasks, such as resource allocation and funding priority setting [38], [75].

This chapter begins with a review of the relevant literature to grasp the current research paradigm and pinpoint its limitations. Drawing from this examination, we identify critical research gaps in areas including *data Use ethics*, *data openness*, *implementation efficiency*, *ease-security trade-off*, and *knowledge complexity and reusability*. The insights garnered from these works and identified gaps lay a solid foundation for subsequent chapters, guiding the development of research data search and management systems, particularly focusing on the GAUDS system.

2.1 Related Work and Challenges

In this section, we analyze the recent literature in Data Search and Reuse, as well as Scholarly Knowledge Graphs for Research Data Management (SKG4RDM). We review the literature in a broad scope from the current landscape of research data practices and needs to the applied domain of using SKGs to enhance RDM. In particular, we review the current states of RDM implementations based on each of the sub-principles in FAIR. We then identify the current advances and limitations of the SKG4RDM implementations. The rest of this section is as follows.

- **Key Goals of Research Data Management**. This section focuses on reviewing the key goals in RDM, namely discoverability and (re)usability.
- **Practices and Need: Research Data Management and Data Citations**. This section discuss research data management from the user perspective, focusing on the current practices and how users discover and reuse data.
- **FAIR Principles and Limitations**. This section introduces the FAIR principles, including Findability, Accessibility, Interoperability, and Reusability. We then discuss the relevant critiques and extensions, focusing on FAIR principles' guidance on research data management.
- **SKG4RDM: Scholarly Knowledge Graphs Implementations for Research Data Management**. In this section, we introduce SKG in the context of RDM. We will also summarize the current advancements in FAIR principles for guiding SKG4RDM, as well as the challenges.
- **Discussions**. This section will summarize the current state of SKG4RDM, identify the limitations in research data management, and map them to the practical needs of data discovery and reuse.

2.1.1 Key Goals: Discoverability and (Re)usability

RDM is essential for ensuring the quality, integrity, and accessibility of research data, thereby fostering reproducibility, transparency, and collaboration in the scientific community [20], [76]. The increasing importance of data-driven research underscores the need for effective RDM strategies, which can help maximize the value of research data and facilitate its reuse in various contexts [77], [78]. Effective RDM strategies encompass a wide range of activities, including data planning, storage, sharing, and preservation [79]. Different disciplines and research contexts may require tailored RDM approaches, leading to variations in the specific practices and priorities emphasized by each field.

Current RDM models are often derived from digital curation models, which focus on the preservation, organization, and accessibility of digital assets [80], [81]. These models emphasize the importance of metadata, data documentation, and adherence to standardized formats and protocols to ensure the long-term preservation and accessibility of research data [82], [83]. Metadata, which provides descriptive, administrative, and structural information about data, plays a crucial role in facilitating discoverability and usability by enabling users to understand the context, content, and relevance of the data [84]. Adhering to community-driven metadata standards and controlled vocabularies can further enhance the interoperability and discoverability of research data [85].

In this section, we overview key goals of discoverability and usability in RDM, as well as related and representative models and standards. Discoverability and usability of research data are key aspects of RDM, and they depend on the implementation of standardized metadata, effective data organization, and adherence to best practices for data sharing and documentation [86].

Discoverability requires that research data are well-documented, indexed, and presented in a way that allows users to come across it through various search mechanisms, browsing, or recommendations [87], [88]. There are several current RDM models and standards that facilitate discoverability of research data. One fundamental way to enhance the discoverability of research data is by using unique and persistent identifiers, such as DOIs (Digital Object Identifiers), which can be used to create stable and citable links to data resources [89]. While there are implementations of search engine optimization (SEO) techniques, which help improve the discoverability of datasets in search results [90], the exact matching of DOIs are more straightforward and accurate than searching-based dataset discovery.

In addition, data repositories, which often implement standardized metadata schemas and controlled vocabularies, can further facilitate the discoverability of research data by enabling effective search and retrieval functionalities [85]. For instance, there are metadata schemas such as the Data Documentation Initiative (DDI) that aims to provide standards for discovering research data in the social sciences domain [91], and data repositories provide standardized interfaces for searching and accessing datasets, such as DataCite [92] and re3data [85]. By adhering to these standardized metadata schemas and utilizing data repositories, researchers can ensure that their datasets are more easily discovered and understood by others in their field. The implementation of these metadata schemas and data repositories also plays a crucial role in addressing the challenges of data heterogeneity, as they help establish a common language and framework for data description and organization [93]. As a result, researchers can more efficiently navigate, locate, and access relevant datasets, thereby accelerating the research process and enhancing the overall impact of research data [78].

(Re)usability refers to the ease with which researchers can understand, process, and utilize well-documented research data for their specific purposes [76], [83], [94].

Enhancing the usability, or reusability[1], of research data can be achieved through proper data documentation, which includes the use of standardized formats and descriptive information about the data, its collection methods, and any processing or transformations applied to it [95]. This not only ensures that researchers can effectively interpret and reuse the data but also supports long-term preservation efforts by providing essential information about the data's context and provenance [83].

Current RDM models for improving usability include the development of domain-specific data standards, such as the CF (Climate and Forecast) Metadata Conventions for climate data [96] and the Darwin Core for biodiversity data [97]. These domain-specific standards help researchers in their respective fields to more effectively understand, integrate, and analyze the data, thereby streamlining the research process for data users. Other approaches involve the creation of user-friendly data repositories with intuitive interfaces [80], [81], which facilitate navigation and data access, ultimately improving user experience and promoting wider adoption of data sharing practices. Moreover, the provision of clear documentation and data dictionaries [77], [78] ensures that researchers can efficiently comprehend the dataset's structure, variables, and metadata, thereby reducing barriers to data use and fostering reusability. Additionally, the use of software tools that facilitate data processing and analysis, such as programming libraries and visualization tools [98], [99], can further enhance the usability of research data by simplifying data manipulation and interpretation, ultimately benefiting a broader range of researchers and stakeholders.

2.1.2 Data Search and Reuse Practices by Providers and Consumers: Research Data Management and Data Citations

There are two main groups of stakeholders in research data search and reuse: (1) the **providers**, mainly including creators, funding agencies, and owners, and (2) the **consumers (users)**, mainly including researchers, scientists, educators, and students [70]. While the role of stakeholders may overlap, for examples, researchers and scientists can also be the creators of datasets, we will use the data provide and consumer perspectives to review practices and need in data search and reuse. In particular, the providers process research data through research data management, and the consumers mark their reuse by data citations.

Research Practices in Data Archives and Impact of Research Data

Many prior studies of data reuse have focused on data reusers' attitudes as shown through interview studies, surveys, and analyses of data requests [100]–[102]. These studies offer insights into researchers' considerations and motivations for seeking data. However, behavioral studies rely on users' accounts of their own data use. From these studies alone, we cannot know how users interact with data throughout the research process, let alone differentiate the types of work data supports.

[1] In this section, we use the words usability and reusability in an interchangeable manner because of the context of data management, which mainly handles previously released datasets for future reference

Citations to other papers reveal the purpose the referenced paper plays in the work that cites it (e.g., background, compare/contrast, motivation) [103]. Similarly, data references reveal how researchers interact with secondary data (i.e., data produced by someone other than the creator) [104]. Researchers rely on both research literature and datasets to support their analysis and writing [105]. Like references to scientific literature, data references "establish evidentiary sources, give credit, and facilitate the discovery and retrieval of materials on which the citing publication is based" [106]. For example, authors might mention features of a well-regarded survey – such as its sampling frame or questions – without analyzing its data. This kind of attribution does not indicate data use and, therefore, can be challenging to detect. The survey and its producers however, deserve recognition for supporting scientific inquiry.

Organizations, such as data repositories, emphasize the importance of properly citing data to give them credit as research objects. Many data providers assign unique and persistent identifiers, such as DOIs, to datasets [107], [108]. Organizations, such as FORCE11, have also convened task forces to propose formal data citation principles [109]. These principles encourage authors to provide full data descriptions, with unique persistent identifiers, in their papers. Some other organization, such as Kaggle, mainly provide datasets for data science and engineering competitions and crowd-sourcing collaborations [110], [111], instead of academic research. While their goals are focusing on enabling direct reuse of datasets, their main purpose is not focusing on enhancing scholarly communication. However, their data curation practices and user-friendly designs can serve as useful references for the data management and frontend design in data search system. For example, the interactive metadata distribution of numerical data fields in datasets can clearly and interactively inform the potential dataset user if a variable fir their data user purpose.

Data archives support data-intensive research by providing long-term data stewardship, access, and high-quality data management. Notable examples of data archives with high levels of curation include GenBank [112], a rich repository of genetic sequence data; SESAR [113], a repository of metadata describing physical samples in the earth sciences, as well as links to derived datasets; and PANGAEA [114], a publisher for georeferenced datasets linked to earth system studies. Data sharing through archives enables researchers to find and reuse data that they did not collect. In other words, data created for one purpose can be used by new audiences to answer new questions [20], [115]. Researchers can use existing data to validate previous findings, extend their data collections, or form the basis for new studies via integration or independent reuse [60], [116]–[118]. Additionally, as more funders and journals mandate that data from grants and papers be shared openly, data archives are only growing in importance as sites of scholarly communication.

The data held in these repositories often have untapped reuse potential across disciplinary boundaries [119], [120]. Such interdisciplinary research using archived data can lead to breakthrough discoveries [78], [121]. Fields of research may share an interest in explaining different aspects of the same phenomenon, giving rise to interfield theories that bridge fields of science [122]. "Borderland disciplines" sometimes form where fields of research collide over shared resources, such as instruments or data, leading to the evolution of new techniques [123]. Datasets that facilitate interactions between research areas therefore function as "boundary objects," carrying multivalent analytical potential across research communities [124] and facilitating knowledge exchange across boundaries. However, there has been little research on the prevalence of such datasets-as-boundary-objects. We know little about which features of datasets promote boundary crossing, or how to measure their collaborative potential.

Networks and Interdisciplinary Communities of Data Consumers

One way of exploring interdisciplinary data reuse – and therefore, the extent to which datasts function as boundary objects between communities – is by studying data citation networks. Efforts to promote data citation over the last 20 years have led to the adoption of new data citation practices in many communities. Milestones formalizing data citation include the Joint Declaration of Data Citation Principles [125], Data Citation Roadmap for Scholarly Data Repositories [126], and Data Citation Roadmap for Scientific Publishers [127]. Data citation counts provide a foundation for studying the scholarly impact of scientific data and the value of data management efforts.

The adoption of data citation principles makes it possible to analyze emerging data reuse behavior and structures of hidden research communities in data citation networks. Citation networks generally represent documents as vertices and citations of one document by another as edges [128]. Citation networks can highlight central nodes like influential institutions; heavy edges between nodes indicate important connections and processes, like the diffusion of ideas [129]. Prior studies of citation networks have provided insights into ties between individual researchers and collaborations between research disciplines [130]. Studies of publication citation networks (e.g., papers or journals) have also identified novel papers, measured the impact of papers and their authors, and attributed discoveries to authors [131].

Whereas publication citations broadly enable lineage retrieval for ideas, data citations indicate the origins and processing history of the datasets that have been used in an analysis [132]. Data citation networks reflect connections between disciplinary literature and the research data that they draw from. They reveal the reach of research data and support the computation of bibliometrics that show the relationships and impacts of scientific products [133]. The interactional context of data production and citation also reflects relationships between data producers and consumers in a broader data economy [70].

The analysis of citation networks can reveal hidden organizational structures. Co-citation analysis studies the structure of science and the emergence of specialities in bibliometric networks by examining how frequently pairs of documents are invoked [134]. Author co-citation analysis reveals individual contributions to speciality areas and paradigm shifts in the research landscape [135]. Citation analysis can be used to identify exclusionary community structures, such as “invisible colleges” [136] – in-groups that control scientific discourse, which are defined by strong ties and informal communication [137]. Similar analyses can also detect “citation cartels” of authors who cite each other exclusively, and effectively shut out other authors who work on the same subject [138]. In addition to exclusionary practices, citation analysis can also identify convergence in research communities. Studies of cross-field citation networks have found that fields of science tend to become more integrated, rather than exclusive, over long periods of time [139], albeit incrementally across neighboring disciplines [140].

While the notion of “community” is central to these analytical methods, it is a difficult concept to operationalize [141]; communities may take many forms, and may play many roles. Identifying communities via data citation is further complicated by the interdisciplinary nature of data analysis and citation [142]. However, we take inspiration from prior work showing that data reuse can be viewed as an indirect form of cooperation and collaboration between researchers – and groups that commonly reuse the same data might be considered communities-at-a-distance [143]–[145]. Research data is a primary input for scientific knowledge production, making data archives important sites for identifying nascent research communities.

Research Data Management Practices and Data Citation Work

Data citation is an emerging and interdisciplinary field of research that studies data in terms of their attribution, connection, discovery, sharing, impact, and reproducibility [71], which are end-user goals for research data management. There are well-proposed standards and infrastructures for creating data citations. A fundamental and well-known standard, the Joint Declaration of Data Citation Principles (JDDCP) encompasses the vitality and legitimacy of data, the need to give scholarly credit to contributors, and the importance of data as evidence [146]. Consequently, adopting the similar approach of scholarly publication citation, current methods of data citation use persistent identifiers (PID), such as the Digital Object Identifier (DOI), for addressing the data citation identification problem [71], [147].

Using the identification methods, research in the past two decades focus on creating data citation systems and infrastructure that can support citing datasets in growing and large-scale scholarly databases with various granularity in different research domains. In 2005, the German National Library of Science and Technology started assigning DOI names to datasets, which led to the funding of DataCite, a global consortium supporting and promoting data citation [148]. The following research of the fundamental work by DataCite advance different aspects of identifying data citations. Some research focus on designing data citation systems that support citing in large-scale and dynamically growing scholarly databases [149], [150]. Some other research focus on user-end applications, such as a web-based system that focuses on the registration of social science data [151], and a data citation generation infrastructure that identifies subsets of datasets based on filtering and sorting methods [152].

Effective dataset retrieval relies upon the quality and quantity of metadata, which provides prospective data reusers with valuable context [153]. Data references offer insights into how researchers make sense of data produced by others (e.g., through re-analysis), and strategies for contextualizing or justifying data choices when communicating to a scientific audience [100], [154]. These insights and strategies provide information that may help science funders, data producers, and data curators, understand how their data are used. For instance, direct or indirect data mentions indicate the proximity between the author and data [155] as well as the type of data reuse (e.g., integrative, comparative) or non-reuse (e.g., discussing an implication, explaining the source of a linked variable) [102], [156].

Prior investigations of data references have restricted their analysis to specific elements of literature (e.g., abstracts) or types of works (e.g., data papers) [155], [157]. Manual approaches to study data references are time and resource-intensive, and do not easily scale [158]. By contrast, automated processes enable tracking and analysis of data references across scientific literature. Thus, mining data references can help capture data's broader impacts and improve the prospect of assigning credit to data producers and providers.

However, even with the development of standards, methods, systems, and infrastructures, formal data citation remains not universal for a variety of reasons. First, data citation practices vary in different fields, where areas of science, engineering, and technology cite datasets frequently, while there are high uncitedness rates in the social sciences and arts and humanities data studies [159]. This uneven uncited rates may be due to the citation practices and focus of research in different fields. Second, the enforcement of data citation from research journals is inadequate. Fewer than a half of journals point toward a style manual for data citation [160]. Moreover, any in-operation data citation standards or systems are originally designed in a non-digital world with limited use cases [161]. Even if there are many established data citation practices, it is impossible to trace the previous citations using these newly created or applied methods. Thus, the automatically extracted bibliographies rarely cover data work before the first use of DOI in data citation. As a result, data users' informal, incomplete, or incorrect practices of citing datasets remain in their research articles, and data citations cannot be retrieved by simple queries.

With methodological advancements in automation methods and learning systems, data citations can be specified and automatically generated for scholarly databases [162]. For examples, based on a set of citation views, automatic citations generating process can create general queries over a relational database [163]. The "learning to cite" can also automatically construct citations to nodes within XML datasets [164]. Using advanced database techniques, including query answering using views and provenance, automated systems can understand and generate complex data citation queries [165]. I particular, the Data Citation Index (DCI) links published research articles to their underlying datasets [166], which has a potential to be be used retroactively in harvesting both formal and informal data citations. The Citation Typing Ontology (CiTO) is another well-adopted standard indicating community agreement on typical elements of a citation[2]. More recently, research suggests use machine learning and natural language processing (NLP) to detect informal data citations [167], and projects like Make Data Count[3] and Show Us the Data[4] also implements machine learning to support data citation discovery.

2.1.3 FAIR Principles and Limitations

In this section, we analyze the FAIR principles, a widely implemented data management guideline, with regard to its RDM use cases and its limitations. The FAIR principles provide guidelines for making digital assets Findable, Accessible, Interoperable, and Reusable [20]. In general, the FAIR principles emphasize machine-actionability because data are increasing in their volume, complexity, and speed [20]. More specifically, the principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure (structures, systems, and resources that enable and support digital objects) across their sub-principles.

Findability refers to the ease with which data can be located by both human users and computer systems. It necessitates the use of machine-readable metadata, which is integral to the automatic identification and discovery of datasets and related services.

Accessibility pertains to the mechanisms through which a user, after locating the necessary data, can gain access to it. The process could entail specific procedures such as authentication and authorization. Interoperability concerns the capacity of data to be integrated with other datasets. Furthermore, the data

[2] https://sparontologies.github.io/cito/current/cito.html
[3] https://makedatacount.org/
[4] https://coleridgeinitiative.org/show-us-the-data/

should be capable of interacting with applications or workflows for various purposes, such as analysis, storage, and processing. Reusability aims to maximize the potential for data reuse. It requires that both metadata and data be comprehensively described to enable their replication or amalgamation in various contexts. The sub-principles include Findable (F1, F2, F3, F4), Accessible (A1, A1.1, A1.2, A2), Interoperable (I1, I2, I3), and Reusable (R1, R1.1, R1.2, R1.3), which will be discussed in detail in the following section.

FAIR principles' workflow and contributions

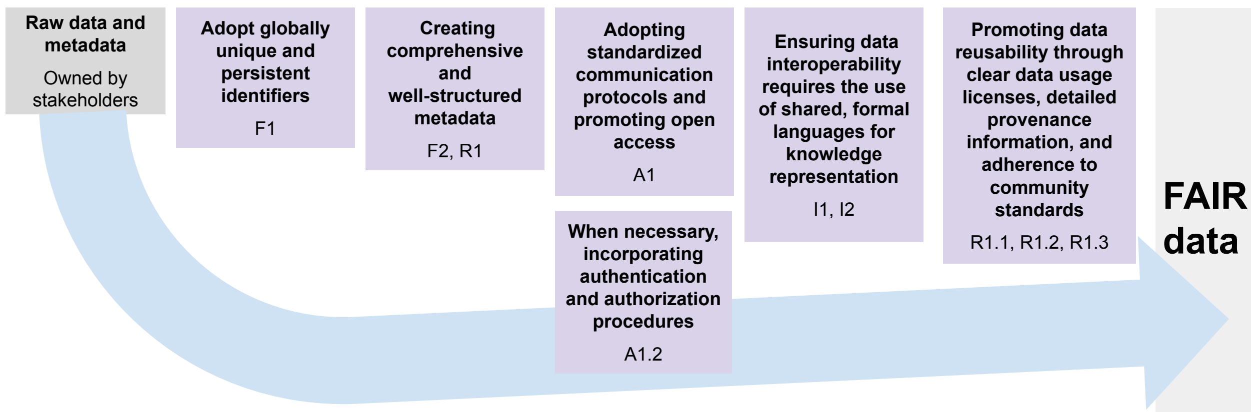

Figure 2.1: A theoretical workflow of implementing the FAIR principles

Applying the FAIR principles to RDM involves several key steps that help ensure data is efficiently managed and utilized within the scientific research community. Figure 2.1 provides a visualized workflow of the following key steps, which cover many actionable sub-concepts in the FAIR principles. First, institutions, researchers, and data repositories need to adopt globally unique and persistent identifiers for datasets and metadata, such as Digital Object Identifiers (DOIs) (F1) [89]. This allows for unambiguous identification and citation of data, contributing to its findability and long-term accessibility. Second, creating comprehensive and well-structured metadata is crucial (F2, R1). Following established metadata standards and schemas ensures consistency and compatibility across datasets [85], [91], [92]. Third, adopting standardized communication protocols and promoting open access to data and metadata fosters accessibility and collaboration in research communities (A1) [76], [78]. When necessary, incorporating authentication and authorization procedures helps protect sensitive data while still maintaining its accessibility to authorized users (A1.2). Moreover, ensuring data interoperability requires the use of shared, formal languages for knowledge representation, such as Resource Description Framework (RDF) [168], and the adoption of controlled vocabularies and ontologies that follow FAIR principles (I1, I2). This facilitates data integration and enables researchers to combine and analyze data from different sources more effectively. Last, promoting data reusability involves providing clear data usage licenses, detailed provenance information, and adherence to community standards (R1.1, R1.2, R1.3) [76], [83]. This enables researchers to understand how to generate and process the data, and under what conditions they can reuse, thereby increasing trust and confidence in the data.

To better demonstrate the FAIR principles' roles in enhancing the key goals of RDM, we analyze each of the sub-principles and their contributions to discoverability and usability (Table 2.1.3). As Table 2.1.3 shows, the FAIR principles substantially contribute to the discoverability of RDM by promoting the use of globally unique and persistent identifiers (F1), which facilitate stable and citable links to data resources. Rich metadata (F2) and registering or indexing (meta)data in searchable resources (F4) further enhance the discoverability of datasets by enabling more effective indexing and searchability. Qualified references to other (meta)data (I3) foster the exploration and discovery of related resources, broadening their visibility within the research community. In terms of usability, the FAIR principles emphasize the importance of clear associations between metadata and data (F3), enabling better understanding and interpretation of the data. Standardized protocols (A1), vocabularies (I2), and domain-relevant community standards (R1.3) facilitate data access, interpretation, and integration, thus improving usability and reusability. Additionally, the principles highlight the significance of maintaining the accessibility of metadata (A2) and providing clear and accessible data usage licenses (R1.1), which together promote data sharing and reuse.

Table 2.1: FAIR principles and its contributions to discoverability and usability

#	**Description**	**Discoverability**	**Usability**
F1	(Meta)data are assigned a globally unique and persistent identifier	Globally unique and persistent identifiers facilitate stable and citable links to data resources, enhancing their discoverability.	Persistent identifiers enable users to consistently locate and access datasets, promoting their long-term usability.
F2	Data are described with rich metadata (defined by R1 below)	Rich metadata allows more effective indexing and searchability of datasets, improving their discoverability.	Comprehensive metadata enables users to better understand the context, content, and structure of datasets, enhancing their usability.

Continued on next page

Table 2.1 – *Continued from previous page*

#	Description	Discoverability	Usability
F3	Metadata clearly and explicitly include the identifier of the data they describe	By explicitly linking metadata to their corresponding data, users can more easily find and explore related resources.	Clear associations between metadata and data enhance the understanding and interpretation of the data, thus improving usability.
F4	(Meta)data are registered or indexed in a searchable resource	Registering or indexing (meta)data in searchable resources increases their visibility and findability within the research community.	By making (meta)data more accessible through searchable resources, users can more readily locate and utilize relevant datasets.
A1	(Meta)data are retrievable by their identifier using a standardized communications protocol	Standardized protocols for retrieving (meta)data by their identifier ensure consistent access to data resources.	By using standardized protocols, researchers can more easily access and work with (meta)data, improving overall usability.
A1.1	The protocol is open, free, and universally implementable	Open, free, and universally implementable protocols promote widespread adoption, increasing the discoverability of (meta)data.	Such protocols ensure that (meta)data are accessible to a broader audience, enhancing their usability.

Continued on next page

Table 2.1 – *Continued from previous page*

#	Description	Discoverability	Usability
A1.2	The protocol allows for an authentication and authorization procedure, where necessary	None.	Authentication and authorization procedures help protect sensitive data while still enabling authorized users to access and work with the data.
A2	Metadata are accessible, even when the data are no longer available	None.	By maintaining the accessibility of metadata even when the data are no longer available, users can continue to learn from and build upon past research.
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	Formal, accessible, shared, and broadly applicable languages for knowledge representation enable better indexing and searchability of (meta)data.	These languages facilitate data integration and harmonization, promoting data interoperability and reusability.
I2	(Meta)data use vocabularies that follow FAIR principles	FAIR-compliant vocabularies enhance the searchability and findability of (meta)data.	Standardized vocabularies improve data interpretation and understanding, fostering usability and reusability.

Continued on next page

Table 2.1 – *Continued from previous page*

#	Description	Discoverability	Usability
I3	(Meta)data include qualified references to other (meta)data	Qualified references to other (meta)data promote the exploration and discovery of related resources.	By linking to related (meta)data, users can more easily navigate and integrate data resources, improving usability and reusability.
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes	Richly described (meta)data with accurate and relevant attributes enable better searchability and findability.	Detailed (meta)data descriptions allow users to more effectively comprehend and utilize datasets, enhancing usability.
R1.1	(Meta)data are released with a clear and accessible data usage license	None.	Clear and accessible data usage licenses facilitate data sharing and reuse by providing users with explicit usage guidelines.
R1.2	(Meta)data are associated with detailed provenance	None.	Associating (meta)data with detailed provenance information allows users to trace the origins and history of the data, enhancing trust and confidence in the data's quality and reliability.

Continued on next page

Table 2.1 – *Continued from previous page*

#	Description	Discoverability	Usability
R1.3	(Meta)data meet domain-relevant community standards	None	By adhering to domain-relevant community standards, researchers can ensure that their (meta)data are more easily understood, integrated, and reused by others within their field.

As such, the FAIR guiding principles [20] provide a framework for the creation of standardized metadata, the use of persistent identifiers, and the development of open and machine-readable data formats, all of which contribute to the improved discoverability and impact of research data in the scientific community [169]. Notability, while the terms "discoverable" and "findable" are often used interchangeably, they have subtle differences in meaning, especially in the context of research data management. Discoverability refers to the extent to which a piece of information or data can be uncovered or encountered, often without an explicit search [93]. It involves making the information or data visible and accessible to users in a way that enables them to stumble upon it. "Findable" data, on the other hand, is well-organized and has high-quality metadata, making it easy for users to locate the data using search queries or filters [76]. In the context of research data, findable data is well-organized and has high-quality metadata, making it easy for users to locate the data using search queries or filters. It regards the use of effective search tools and strategies as an external process, while focusing on clear organization and categorization of the information itself.

Similarly to their role in improving discoverability, the FAIR guiding principles [20] provide a framework for researchers to create reusable datasets that are more readily accessible, both through open platforms and with clear usage guidelines. The focus on Interoperability ensures that datasets are structured and formatted in a way that allows for seamless integration with other datasets, enabling researchers to collaborate and build upon existing knowledge more efficiently. The Reusability aspect of the FAIR principles encourages the provision of comprehensive documentation and metadata, which promotes a clear understanding of the data, reduces barriers to its use, and facilitates diverse applications across various research domains [169]. However, it is important to notice that there are substantial differences between the general usability goal of RDM and the reusability principle in FAIR. Regarding the motivation of creating and maintaining research data, usability means that the data is well-documented, easy to understand, and can be readily used for its intended purpose. Reusability, here, refers to the extent to which a piece of information or data can be utilized for purposes other than its original intention through machine-readable and actionable formats. It is seemingly straightforward that usability focuses more on its original use case and reusability emphasizes on the additional value of research data other than, while usability has a stronger focus on the comprehensive workflow of utilizing research data. It focuses on the "users:" usability of research data enables researchers to create and share data in manners that are more user-friendly, fostering an effective and collaborative research ecosystem.

Limitations of FAIR principles and potential problems

Despite their popularity, the FAIR principles have been subject to critiques, and previous researchers have proposed extensions, which highlight potential limitations and areas for improvement. There are three main limitations identified in nine selected publications, including data use ethics, data openness, and implementation efficiency for users. Table 2.2 summarizes limitations with representative references.

Table 2.2: Critiques and extensions of FAIR principles

Limitations	Related papers (with category and link)	Corresponding key points
Data use ethics (data are findable, but stakeholders are not)	[25] (Critique)	FAIR principles reinforce existing power dynamics and inequalities in data sharing.
	[19] (Critique)	FAIR principles are lacking in ethical data sharing and governance.
	[72] (Extension)	Authors emphasize the importance of respecting indigenous rights and interests in data governance and extend to CARE principles (Collective Benefit, Authority to Control, Responsibility, and Ethics).
Data openness (restrictions on data access)	[170] (Critique)	FAIR data is not equal to open data. FAIR principles focus on data management and discoverability, open data emphasizes unrestricted access and use.
	[169] (Critique)	FAIR implementation can be unsuccessful due to the costs and complexity of data management.
Implementation efficiency (cost, time, and resource needs due to complexity in data and the principles)	[26] (Extension)	FAIR implementation can be unsuccessful due to the costs and complexity of data management.
	[171] (Extension)	FAIR principles are aspirational but not operable. There is a need for an ontological representation of FAIR principles for data practitioners.

Regarding data use ethics, [25] questions the fairness of the FAIR principles, arguing that they may inadvertently reinforce existing power dynamics and inequalities in data sharing. The authors contend that the principles lack explicit consideration of ethical and social implications, such as data ownership, consent, and privacy. They suggest that the principles may not be sufficient to address the complexities of sensitive data, particularly in the context of indigenous and marginalized communities. There are also limitations of the FAIR principles on the lack of guidance on ethical considerations and the potential for misinterpretation of the principles [19]. The authors suggest that the principles should be expanded to cover ethical data sharing and governance, particularly regarding data access and reuse.

There are also practical implementations that extend and improve the FAIR principles in domain-specific scenarios. [72] proposes integrating the CARE principles with the FAIR principles to address the unique challenges of managing indigenous data. The authors emphasize the importance of respecting indigenous rights and interests in data governance, which extends the ethical considerations of the FAIR principles. [172] introduces the concept of 'methodological data fairness', extending the principles of FAIR by considering fairness in the use and analysis of data. In response to the limitations mentioned above, the authors argue for the need to avoid biases and ethical issues in health-related social media research. In data products, not only data itself should be discoverable, but who owns and makes the data should also be recognized. Both of these practical extensions of FAIR principles ensure high-quality data use and aim to mitigate misuse.

Regarding data openness, recent work compares the FAIR principles with the concept of open data. In [170], the authors highlight the differences and synergies between the two approaches, noting that while the FAIR principles focus on data management and discoverability, open data emphasize unrestricted access and use. They suggested that both approaches can contribute to improving data sharing and reuse. In addition, more research investigates the relationship between the FAIR principles and data openness further, elucidating that FAIR is not synonymous with 'open' [171]. The authors further introduce the idea of ontological representation to support the effective operationalization of the FAIR principles. They argue that semantic technologies such as ontologies can improve data interoperability and reusability, which are key aspects of the FAIR principles. In line with this, the authors propose a blueprint for creating 'FAIRer' data sources, including the development of a generic FAIR ontology that can be tailored to specific research domains. Such a FAIR-compliant approach, while not necessarily implying openness, can help to promote open science.

Regarding implementation efficiency for users, recent research investigate the adoption of FAIR principles in the pharmaceutical industry [26]. The study identifies barriers to FAIR implementation, including the costs and complexity of data management. The authors highlight the need for industry-specific guidance and support to facilitate FAIR implementation. Moreover, the FAIR principles are not directly operational [169]. Taking the European Open Science Cloud (EOSC) as an example, the authors argue that for the FAIR principles to be effectively applied, they must be incorporated into the workflows and policies of research institutions and data repositories. They propose the use of machine-actionable data management plans (maDMPs) as a way to operationalize FAIR.

There are specific potential limitations of the FAIR principles regarding applications in the RDM domain. First, while the FAIR principles emphasize the importance of making data findable through persistent identifiers, metadata, and indexing, they do not explicitly address issues of data ownership nor making the constraints of consent and other ethical considerations explicit, which may affect data discoverability. Second, the accessibility principle advocates standard protocols and authentication mechanisms. However, it does not provide detailed guidance on how to share data from the users' perspective, including the necessary technology and prior knowledge needed, which may affect access to large scale or complex data. Third, the interoperability principle calls for data usability, general machine readability and actionability. However, challenges of metadata standards, data quality, and industry-specific requirements may hinder interoperability. Finally, the reusability principle emphasizes data usability through clear licensing, provenance information, and data documentation. The limited emphasis on reusability does not adequately address issues of credit and recognition for data contributors, nor does it explicitly consider fairness of ability and capacity in data use and analysis. Thus, the FAIR principles are insufficiently specified to inform RDM.

2.1.4 SKG4RDM: Scholarly Knowledge Graph Implementations for Research Data Management

Scholarly Knowledge Graphs (SKGs) are large-scale efforts to connect research persons and objects. They leverage comprehensive frameworks to encapsulate the structure, dynamics, and evolution in scholarly ecosystems [50], [51]. In SKGs, academic entities and their interconnections are represented as nodes and edges in a graph. These graphs are composed of a myriad of elements, including authors, articles, journals, disciplines, institutions, data, and the relationships among them, allowing for a multilayered and nuanced understanding of the academic landscape [173]. By providing a holistic view of academic intellectual output, the SKG enhances knowledge discovery and dissemination, fosters interdisciplinary research, and enables innovative methods to measure research impact beyond traditional citation metrics [174]. In addition, it empowers researchers with tools to trace the genealogy of ideas, visualize research trends, identify gaps in the literature, and find potential researchers and objects, such as collaborators with complementary expertise and data specific to a research purpose [175]. Despite its significant benefits, the construction and maintenance of the SKG pose substantial challenges, including the need for effective information extraction techniques, ontology management, and graph analytics, as well as addressing issues related to data quality, interoperability, and privacy [176].

The connections among research persons and objects in SKGs are crucial for managing research data and metadata. For instance, one prominent graph in life and physical sciences connects over 2,200 datasets with half a million computer science and biology publications [177]. There has also been growing interest in linking survey data and related variables to improve data discovery in social sciences [178]. Establishing explicit links between variables, study-level data, and research publications supports the analysis of bibliometric trends and the creation of research metrics to understand the role of data in scientific knowledge production [179]. Acknowledging links among stakeholders behind data and regarding the linked metadata as knowledge infrastructure further contributes to core RDM needs, such as identifying the data lifecycle, gaining a clearer understanding of the stakeholders behind data, and harnessing advanced technologies for automated data annotation, discovery, and linking [180], [181].

Besides expanding metadata for connected research products and formalizing citation links in SKGs, various computational methods help capture the context in which research data are used. These methods assist users in dealing with information overload by offering a summary-level view of knowledge production contexts. Techniques such as citation mining and context extraction promote "context-driven discoverability," which enriches data descriptions based on the disciplines and topics that have reused the data, supplementing the original metadata provided by the data creator and curators upon publication [182]. Similarly, text summarization methods like the TLDR model [183] and knowledge base creation [184] enable citation context modeling and extraction. These methods deduce the main contributions of research publications by extracting relationships from scientific articles, and can be applied as add-ons to large-scale SKG implementations to increase dataset discoverability and usability.

In this regard, SKGs and relevant techniques can offer a structured and interconnected representation of research data, metadata, and provenance information, thus improving data findability, accessibility, interoperability, and reusability (FAIRness) as mandated by the FAIR principles. Previous works in SKGs and RDM have applied FAIRness as a combination of guiding principles and evaluation criteria: [185] conceptualize and implement research data objects' assessment through the FAIR principles; [28] generate literature surveys using SKGs as the infrastructure and FAIR as the guidelines; [186] apply FAIR principles to develop the backbone of European Open Science Cloud (EOSC) across three research categories, including scholarly information, research data, and research software.

In this study, we build on the approach of measuring the FAIRness of data, systems, and infrastructures, using each of the sub-principles in FAIR to evaluate SKG4RDM implementations. We analyze 41 selected RDM implementations through SKGs across fields of research and application focuses. We use three levels, namely Yes, Partially, and No, to assess the compliance of the SKG4RDM implementation with respect to how they implement each of the FAIR principles. Here, we define the annotation criteria as follows:

- Yes: Fully complying (compliance score = 2). The function of the sub-principle of FAIR is sufficiently addressed or implemented as an automated feature in the SKG4RDM implementation, with or without mentioning the name of the principle.
- Partially: Somewhat complying (compliance score = 1). The function of the sub-principle of FAIR is mentioned in the SKG4RDM implementation, while essential human efforts are needed. The human efforts vary from obtaining data access through institutional subscriptions to manual look up metadata schema, depending on the specific sub-principle.
- No: Not complying (compliance score = 0). The function of the sub-principle of FAIR is not mentioned in the SKG4RDM implementation.

Table 2.3 shows the summary statistics of the overall compliance per sub-principle and the variance among individual SKG4RDM implementations. Overall, there is a high-compliance of the FAIR principles in the current SKG4RDM implementations, and the variance among different individual cases is also generally low. Among all 13 FAIR sub-principles (of different levels), I3 ((Meta)data include qualified references to other (meta)data) as the highest compliance score of 2 with a variance of 0, indicating that all 41 implementations closely follow the I3 principle of internal (meta)data connectivity (marked with green color font in Table 2.3). At the same time, there are two sub-principles with generally low compliance scores, namely A1.2

(The protocol allows for an authentication and authorization procedure, where necessary) and R1.1 (R1.1: (Meta)data are released with a clear and accessible data usage license), showing the complexity of implementing FAIR SKG4RDM applications through the current understanding of opening up data and science, as well as practices in data licensing (marked with red color font in Table 2.3). For sub-principle A1.1 (The protocol is open, free, and universally implementable), while the median compliance score is not low, the variance of compliance is the highest among all implementations, suggesting the existence of different understanding and practices of research data openness and other accessibility measurements.

Table 2.3: Summary statistics of FAIR-compliance scores of 41 SKG4RDM implementations

Summary Statistics	**F1**	**F2**	**F3**	**F4**	**A1**	**A1.1**	**A1.2**	**A2**	**I1**	**I2**	**I3**	**R1.1**	**R1.1**	**R1.2**	**R1.3**
Median	2	2	2	2	2	2	1	2	2	2	2	2	1	2	2
Variance	0.16	0.16	0.02	0.24	0.24	0.54	0.36	0.4	0.02	0.16	0	0.13	0.37	0.05	0.14

The rest of this section reviews SKG4RDM implementations using the FAIR principles. We will first overview the compliance (compliance rate, i.e., the proportion of the level of compliance among all 41 implementations) of the parent-level FAIR principles. We will then summarize the identified patterns of using popular frameworks or tools to implement SKG4RDM applications. Finally, we assess the FAIRness with regard to each of the sub-principles and provide representative examples. We will also provide critical analysis of SKG4RDM implementations regarding the FAIR principles. In addition to checking how well the FAIR principles are met in each implementation, we evaluate from the user's perspective and point out limitations.

Findability

Figure 2.2 shows the assessment of the current implementations of the Findability principle. We observe an overall high compliance of the current SKG4RDM implementations with regard to this principle, with an average of 79.27% full-compliance rate, an average of 20.73% partially-compliance rate, and without non-compliance implementations.

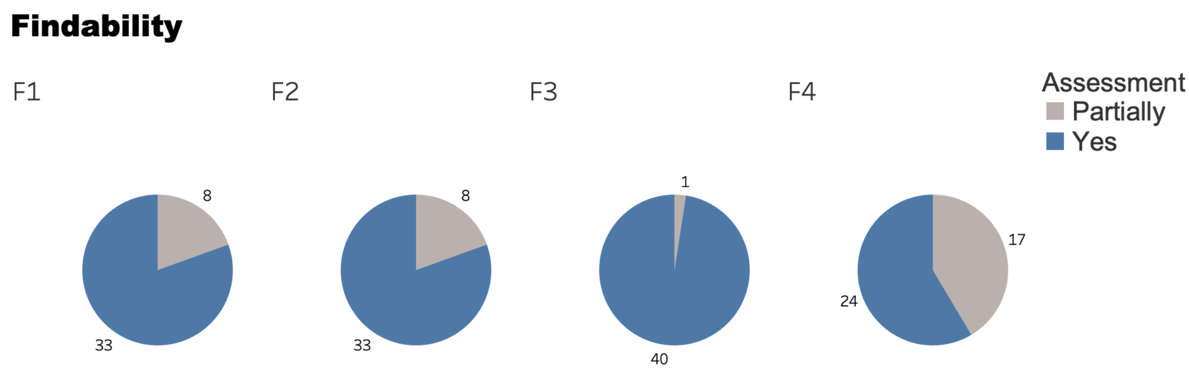

Figure 2.2: Assessment of findability in current SKG4RDM applications

F1: (Meta)data are assigned a globally unique and persistent identifier The F1 principle ensures RDM assigns globally unique and persistent identifiers to every element of metadata and every concept or measuret in a dataset. These identifiers often consist of an internet link, ensuring that they are easily accessible and resolvable, removing ambiguity in the meaning of published data. Some general-purpose examples of such identifiers include the Digital Object Identifiers (DOIs) [187], as we mentioned in previous sections as one of the most famous and widely used globally unique and persistent identifiers. Other general-purpose examples include Identifiers.org that provides resolvable identifiers [188] and Universally Unique Identifiers (UUIDs) that ensure uniqueness through a combination of time and spatial analysis [189] . These identifiers are applicable to almost any type of research objects and persons. At the same time, there are user-specific identifiers for RDM stakeholders, such as funders, institutions, and researchers, through resources like CrossRef Funder Registry, Research Organization Registry (ROR), and Open Researcher and Contributor IDentifier (ORCID).

The current state of compliance with F1 varies across different datasets and domains, with some adhering to these principles while others only partially complying. Either depending on the above universal and unique identifiers or creating their own, many current SKG4RDM applications carefully adhere to the Findability principle and increase research datasets' discoverability. There are several well-known SKG4RDM applications that adhere to F1 well. The Microsoft Academic Graph (MAG) assigns globally unique and persistent identifiers to the metadata [62]. Open Research Knowledge Graph (ORKG) uses globally unique and persistent identifiers (e.g., DOI and ORCID) for resources [181]. AIDA, a knowledge graph about research dynamics in academia and industry, assigns globally unique and persistent identifiers to its entities [27].

At the same time, among the 41 SKG4RDM implementations, eight (19.51%) are only partially adhering to F1. problems of using either non-unique or non-universal identifiers. A small proportion of the lack of uniform adoption of these unique and persistent identifiers can result in the need for applying disambiguation techniques on all potential datasets, which can cause harm to research data's discoverability and usability. For example, KGMM, a maturity model for SKGs based on intertwined human-machine collaboration, discusses the use of unique and persistent identifiers, but it is not clear if they are implemented across all data and metadata [190]. The less complete implementations of the F1 principle, like KGMM, can make it difficult for users to locate specific datasets and can potentially hinder data sharing and collaboration efforts.

F2: Data are described with rich metadata The F2 principle emphasizes the importance of providing extensive and detailed metadata for RDM. Self-describing metadata ensures that data can be discovered and understood based on the provided information, even without the data's identifier. This includes both intrinsic metadata, which is automatically captured by machines generating data (e.g., DICOM information for image files [191]), and contextual metadata, which covers elements such as the protocol used, measurement devices, units, species involved, and various other aspects of the study. Providing comprehensive metadata enables computers to automate routine sorting and prioritizing tasks, which would otherwise demand significant effort from researchers.

Several frameworks and tools have been developed to support rich metadata creation and management. The Dutch Techcentre for Life Sciences (DTL) metadata editor is a user-friendly tool that aids researchers in generating and editing metadata records. The Data Catalog Vocabulary (DCAT) framework provides a standard model for describing datasets in data catalogs, which assists in the process of marking up datasets with rich metadata. The ISA (Investigation, Study, Assay) framework offers a structured, hierarchical approach to organizing and describing metadata that helps in capturing and managing various types of research data. These frameworks and tools not only promote metadata best practices but also ensure interoperability and consistency across different research domains.

While many datasets have adopted rich metadata using tools similar to the above, the compliance of F2 is not universal, and there is no clear definition of what constitutes "enough" metadata. This lack of a standardized benchmark can make it difficult to compare datasets or ensure consistent quality across different data sources. Among the 41 literature, 8 of them (19.51%) are less sufficiently adhering to F2. For example, [192] provides metadata, but it is not clear whether it is rich enough to meet this criterion. In [33], the metadata is provided, but it may not be as rich as desired for all use cases. Moveoever, in [29], the paper provides rich metadata for some entities, but there is room for improvement in describing all entities in a more comprehensive manner. In essence, rich metadata can sometimes seem more like a slogan than a well-defined standard, which can create confusion and make it challenging to enforce best practices. The benchmarking and crowdsourcing efforts, such as the "rich content competition" that discover relationships and build new metrics to describe data use, can create standards to ensure metadata richness.

F3: Metadata clearly and explicitly include the identifier of the data they describe The F3 principle emphasizes that metadata should clearly and explicitly include the identifier of the data they describe. In RDM, metadata and the dataset they describe are typically separate files. To establish a strong association between a metadata file and its corresponding dataset, the dataset's globally unique and persistent identifier should be explicitly mentioned in the metadata. This can be achieved by incorporating a schema that explicitly links the metadata to the dataset. For example, the OpenRefine Metadata Extension enables users to create, edit, and export metadata in various formats while maintaining the explicit linkage between metadata and the datasets they describe [193]. By utilizing tools like the OpenRefine Metadata Extension, researchers can ensure their metadata adheres to the F3 principle, easily locating, accessing, and understanding the connection between metadata and the data they describe, facilitating data discoverability and reuse.

In general, compliance with F3 is high (58.53%, 24/41), as including the identifier in the metadata is relatively straightforward. For example, the enhanced Microsoft Academic Knowledge Graph (MAG) includes the identifiers of the data they describe, which allows for easy linking and discovery [194]. AceKG, a large-scale SKG designed for academic data mining, clearly and explicitly includes the identifier of the described data [195]. Only one implementation is less sufficiently adhering to F3. [30] presents a method for extracting knowledge graphs from metadata, but it does not explicitly discuss the retrieval of metadata by their identifier using a standardized communications protocol.

F4: (Meta)data are registered or indexed in a searchable resource The F4 principle highlights the importance of registering or indexing (meta)data in a searchable resource for effective RDM. Simply having well-structured and documented (meta)data is not sufficient; it must also be easily discoverable by search engines and other data retrieval services. By ensuring (meta)data are registered or indexed in searchable resources, researchers can improve data discoverability, accessibility, and the potential for reuse and collaboration, thereby enhancing the overall impact of their research. An example tool that supports the F4 principle is the FAIR Data Point. The FAIR Data Point is an open-source software solution that allows researchers to publish their (meta)data in a standardized and searchable manner. By using this tool, researchers can create FAIR Data Points for their datasets, making them easily discoverable and accessible through search engines and other data retrieval services. By leveraging such tools, users can ensure their (meta)data adhere to the F4 principle, ultimately contributing to the advancement of FAIRness.

Regarding the current SKG4RDM implementations, 41.46% of the datasets (17/41) only partially comply with F4, as there is no unified indexing or registration system for all types of data. This can make it difficult for users to locate specific datasets, even if they are described with rich metadata and assigned unique and persistent identifiers. The lack of a standardized system for registration and indexing means that users may need to search through multiple platforms to find the data they are looking for. This can be time-consuming and frustrating, and it hinders the efficient sharing and reuse of data across different domains. For example, in the construction and application of materials knowledge graph based on disambiguation, the authors construct the SKG while it is not clear if the (meta)data are registered or indexed in a searchable resource [196]. Similar lack of registration and indexing also happens in the ontology and NLP-based automatic SKG [197] and the Web of Scholars implementation [198], which make those applications internal tools rather than open science assistance that improves general research data discoverability and usability.

Efforts to improve the current state of F4 compliance should focus on developing unified indexing and registration systems that cater to various types of data, making it easier for users to find and access the information they need. Additionally, promoting the adoption of these systems and encouraging researchers to register their datasets will further improve the overall findability of data in the research community.

Accessibility

Figure 2.3 shows the assessment of the current implementations of the Accessibility principle. We observe an overall less than sufficient compliance of the current SKG4RDM implementations, with an average of 49.39% full-compliance rate, an average of 35.98% partial-compliance rate, and an average of 14.63% non-compliance rate.

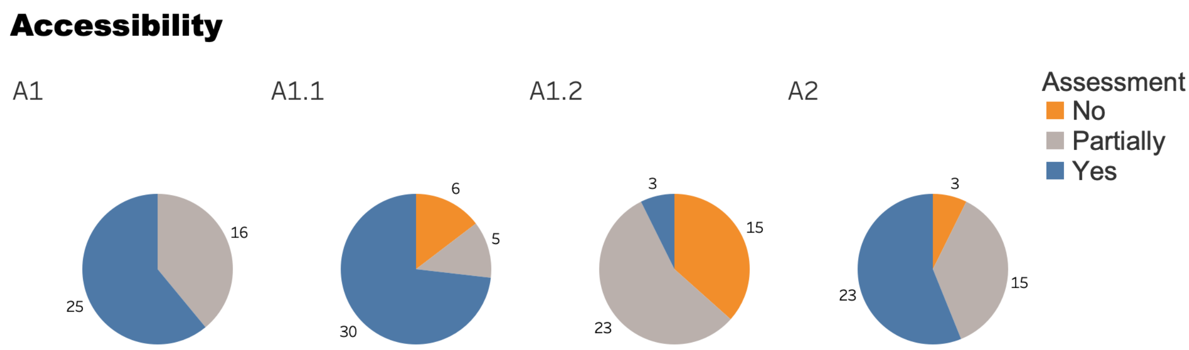

Figure 2.3: Assessment of accessibility in current SKG4RDM applications

A1: (Meta)data are retrievable by their identifier using a standardized communications protocol The A1 principle emphasizes that (meta)data should be retrievable by their identifier using a standardized communications protocol, especially in the context of RDM. Internet users typically retrieve data by clicking on a link, which initiates a low-level protocol, namely Transmission Control Protocol (TCP), to load data in the user's web browser [199]. To adhere to the FAIR principles, data retrieval should not rely on specialized or proprietary tools or communication methods, but standardized communications protocols similar to the functionality of TCP, ensuring that data remains accessible to a wide range of users.

By adhering to the A1 principle, researchers contribute to the advancement of FAIR research data management practices, promoting data accessibility and reusability. For instance, the RMap Project captures and preserves associations among publications and allows retrieval of metadata by their identifier using HTTP [36]. In an application of crowdsourcing scholarly discourse annotations, metadata are retrievable using the hypothes.is API, the application programming interface [35]. In an RDF-based SKG from early modern history, metadata is retrievable by its identifier using a database language SPARQL [200].

While the majority of the SKGs implementations comply with the Accessibility principle, there are some implementations (16/41, 39.02%) that only partially comply with A1, often due to restrictions or the fact that not all data are open source. For example, Twitter data and some institutional data may be limited in accessibility. Additionally, data sharing can sometimes be the last step in a research project, taking considerable time to become available. This can lead to inconsistencies in data accessibility and hinder the overall compliance with A1.

A1.1: The protocol is open, free, and universally implementable The A1.1 principle emphasizes that the communication protocol used for retrieving (meta)data should be open, free, and universally implementable in the context of RDM. By adopting a protocol that is free of cost and open-sourced, researchers can ensure that their data is easily accessible and retrievable by anyone with a computer and an internet connection. This principle promotes the FAIR guidelines by facilitating data reuse and maximizing the impact of research data. Adherence to the A1.1 principle also impacts the choice of the repository used for sharing research data. Researchers should opt for repositories that support open, free, and universally implementable protocols to ensure their data remains accessible and in line with FAIR principles.

There is a small number of implementations that do not comply with A1.1 or only partially comply. The lack of incentives for researchers to adopt open, free, and universally implementable protocols contributes to this problem, as it might not be practical for them to make their data accessible in this manner. Such implementations include AIDA [27], which does not explicitly mention what the protocol is or if the protocol is open and free, and an SKG from survey article tables [28] is also lacking details about how to retrieve its data and metadata.

A1.2: The protocol allows for an authentication and authorization procedure, where necessary The A1.2 principle in research data management (RDM) highlights the importance of incorporating authentication and authorization procedures within the communication protocol when accessing (meta)data. This principle acknowledges that not all data should necessarily be ’open’ or ’free’ and that data access conditions should be clearly specified. Ideally, these specifications should be machine-readable, enabling automated understanding and execution of access requirements or user alerts. Authentication and authorization procedures are essential for ensuring secure access to sensitive or restricted data. These procedures can also be used to authenticate dataset owners or contributors and set user-specific rights. The choice of a repository for sharing research data should take this criterion into account, as it may require user account creation or other access controls.

A1.2 has the largest amount of non-compliance and partial-compliance across all sub-principles. Organizing authentication and authorization procedures can be challenging, as not every data provider is an institution with the resources to manage access requirements effectively. For example, the authentication and authorization procedure is not discussed in the [29] paper about SKG construction and the principle is not mentioned in the [30] paper about SKG enhancement. The lack of such discussion is often due to the researcher’s recognition of their scope.

A2: Metadata are accessible, even when the data are no longer available The A2 principle emphasizes the importance of maintaining metadata accessibility even when the associated data is no longer available. Over time, datasets may degrade or disappear due to the costs of sustaining an online presence for data resources. When this occurs, links can become invalid, leading to wasted time and effort as users search for data that might no longer exist. As metadata storage is generally easier and more cost-effective than data storage, the A2 principle advocates for the persistence of metadata regardless of the data’s availability. This helps ensure that users can still access essential information about the data even if the data itself is no longer accessible. The A2 principle is closely related to the registration and indexing issues described in the F4 principle, as both aim to enhance the discoverability and accessibility of research data and metadata.

An example of a resource that adheres to the A2 principle is the Microsoft Academic Graph (MAG), a large-scale knowledge base with interconnected entities such as academic publications, authors, and institutions [173]. MAG maintained metadata records even if the underlying data is no longer available, ensuring that users can access crucial information about the research data and its context. When the service of MAG ended in December 31, 2021, its metadata can still be accessible and used by other services. For example, Open Alex, a successor of MAG utilizes MAG's metadata.

There is still a significant amount of non-compliance and partial compliance with A2, as the knowledge infrastructure ideologies are not widely applied. For example, [201] generated SKGs within the scholarly domain, while there is no mention of metadata availability when the data is no longer accessible. Similarly, [29] constructed a fine-grained SKG without explicitly mentioning metadata availability. Ensuring that metadata remains accessible even when the data are no longer available can help users avoid wasting time searching for unavailable resources and improve overall data accessibility. Efforts should be made to promote the adoption of A.2 and encourage the preservation of metadata even in the absence of the corresponding datasets.

Interoperability

Figure 2.4 shows the assessment of the current implementations of the interoperability principle. We observe an overall high-level compliance of the current SKG4RDM implementations, with an average of 92.68% full-compliance rate and an average of 7.32% partial-compliance rate.

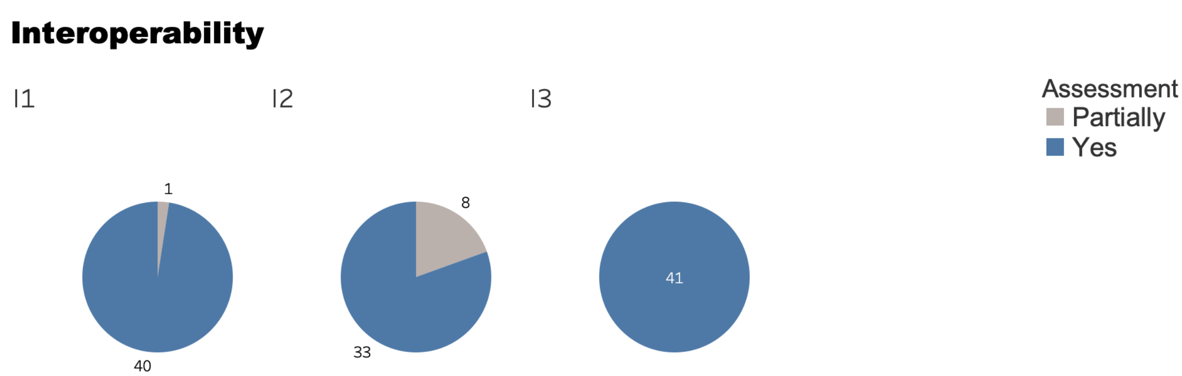

Figure 2.4: Assessment of interoperability in current SKG4RDM applications

I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation The I1 principle focuses on ensuring that metadata uses a formal, accessible, shared, and broadly applicable language for knowledge representation. This principle aims to facilitate a common understanding of digital objects by employing a language capable of representing these objects in a universally interpretable manner. To adhere to the I1 principle, the chosen knowledge representation language should have a formal specification, meaning that its syntax and grammar are precisely defined. Additionally, the language's specifications should be shared and accessible, allowing others to learn and use the language. Lastly, to promote interoperability, the language should be designed for use in various scenarios.

One of the most well-known languages in RDM is RDF (Resource Description Framework), as we mentioned in previous sections, an extensible knowledge representation model used to describe and structure datasets, with the Dublin Core Schema as a notable instance. Benefiting from RDF and related universal languages suitable for networks, compliance with the I1 principle is generally high. For instance, [202] unveils scholarly communities uses RDF-based SKGs, and [203] uses RDF to extract linked open data in biodiversity science and create a SKG, OpenBiodiv, based on these extracted information.

I2: (Meta)data use vocabularies that follow FAIR principles The I2 principle emphasizes that metadata should use vocabularies that follow FAIR (Findable, Accessible, Interoperable, and Reusable) principles. This means that the controlled vocabulary employed to describe datasets must be documented and resolvable using globally unique and persistent identifiers. At a minimum, the vocabulary and its terms or concepts should adhere to the F1, A1, and I1 principles. This means that the vocabulary should have globally unique and persistent identifiers (F1), be resolvable using a standardized communication protocol (A1), and be described with a formal, accessible, shared, and broadly applicable language for knowledge representation (I1).

One general example of ensuring compliance with the I2 principle is through the use of the FAIR Data Point, as we introduced. By adopting this approach, researchers and data managers can ensure that the metadata vocabularies they use meet the necessary FAIR criteria, thus promoting more efficient and effective sharing, understanding, and reuse of research data across various domains. In RDM, specifically, most applications also adhere to FAIR vocabularies, with exceptions due to specific research purposes. For example, [35] targets crowdsourcing scholarly discourse annotations and need to adapt to variations of describing research data. [34] focused on qualitative analysis and had to reuse domain-specific existing vocabularies and create custom ones. While these implementations are well-servicing their own research, it is recommended that they comply with I2 to increase general discoverability and usability of datasets in their research.

I3: (Meta)data include qualified references to other (meta)data The I3 principle highlights the importance of including qualified references to other metadata or data within metadata. A qualified reference is a cross-reference that clarifies its intent, providing essential context for users to understand the relationship between datasets. The goal is to create as many meaningful links as possible between metadata resources to enrich the contextual knowledge about the data while balancing the time and energy required to develop a robust data model.

To achieve this, researchers should specify if one dataset builds upon another dataset, if additional datasets are needed to complete the data, or if complementary information is stored in a separate dataset. In particular, the scientific links between datasets must be described to ensure proper understanding of the relationships between them. Furthermore, all datasets need to be properly cited, including their globally unique and persistent identifiers, to facilitate accurate referencing and attribution.

Compliance with the I3 principle is high, with most datasets including qualified references to other (meta)data. For instance, [204]'s demonstration of faceted search on SKGs maintains qualified references to other (meta)data using RDF triples. By providing more explicit and detailed relationships between datasets, the I3 principle promotes better understanding, discoverability, and reuse of research data within the broader scientific community.

Reusability

Figure 5 shows the assessment of the current implementations of the Interoperability principle. We observe an overall less than sufficient compliance of the current SKG4RDM implementations, with an average of 49.39% full-compliance rate, an average of 35.98% partial-compliance rate, and an average of 14.63% non-compliance rate.

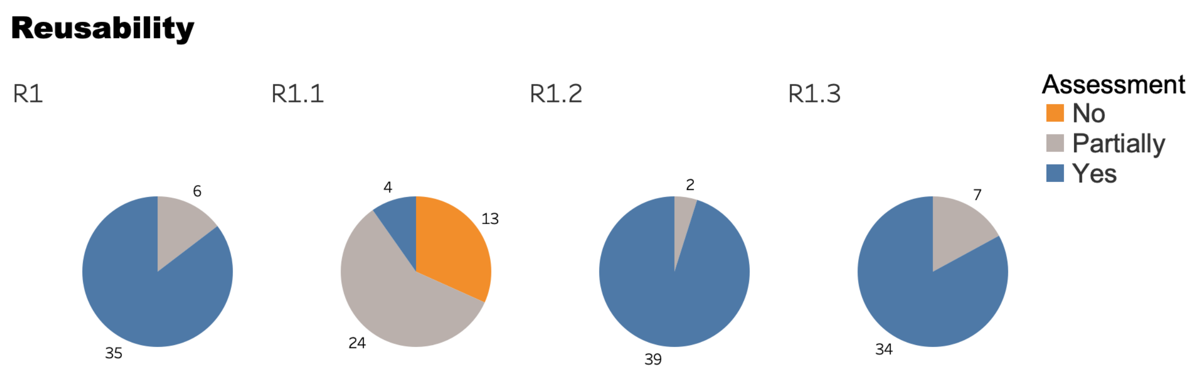

Figure 2.5: Assessment of reusability in current SKG4RDM applications

R1: (Meta)data are richly described with a plurality of accurate and relevant attributes The R1 principle emphasizes the importance of richly describing metadata with a plurality of accurate and relevant attributes. This principle aims to provide comprehensive context for the data, including experimental protocols, the manufacturer and brand of equipment or sensors, species used, drug regimes, and more. While R1 is related to F2, its focus is on enabling users, whether machines or humans, to determine if the data is useful within a specific context. Moreover, R1 encourages data publishers not to attempt to predict the identity and needs of data consumers. The term "plurality" reflects the recommendation for metadata authors to be as generous as possible in providing metadata, even including information that might seem irrelevant at first glance.

To comply with the R1 principle in RDM, researchers should describe the scope of their data, including the purpose for which it was generated or collected. They should mention any particularities or limitations about the data that other users should be aware of and provide details such as the date of generation or collection, lab conditions, data preparers, parameter settings, and software names and versions. Additionally, researchers should clarify whether the data is raw or processed, ensure that all variable names are explained or self-explanatory, and clearly specify and document the version of the archived or reused data.

The R1 principle is largely implemented in current SKG4RDM implementations, with only a few cases of partial compliance. However, R1 does not actively promote usability, and improvements could be made to encourage more comprehensive metadata provision. For example, the open research knowledge graph for CrowdRE maintains richly described metadata [37], but there is room for improvement in providing a plurality of accurate and relevant attributes for all entities. The similar potential improvements should also be applied to the RMap project [36], the Trellisnet-CRF use case [205], and the Canton Canon Digital Library use case [32], which need metadata description that are more detailed and structured, rather than only being explicitly on a subset of metadata.

R1.1: (Meta)data are released with a clear and accessible data usage license The R1.1 principle focuses on the importance of releasing metadata with a clear and accessible data usage license, ensuring legal interoperability. This principle emphasizes that the usage rights attached to the data must be explicitly described to avoid ambiguity, as unclear licensing restrictions can significantly limit the reuse of data by organizations striving to comply with these restrictions. As automated searches become more prevalent, incorporating licensing considerations, clarity of licensing status will grow increasingly important for both machines and humans.

To adhere to the R1.1 principle, researchers should choose well-known and widely accepted licenses, such as the MIT License or Creative Commons licenses, and link them to their data. By employing commonly used licenses, researchers can facilitate a greater understanding of the conditions under which their data can be used, promoting increased data sharing and collaboration within the scientific community. By ensuring that both machines and humans can easily comprehend the licensing status of their data, researchers contribute to the broader goals of transparency and accessibility in RDM.

Many datasets only partially comply with R1.1 or do not comply at all. Licensing information is often omitted from publications, or it is included on repositories like GitHub [29], [201], [206], which are not explicitly mentioned in the publication. Some licenses are organization-specific rather than data product-specific, thus not mentioned in publications, complicating tracking and compliance [31], [32]. Interestingly, R1.1 is a seemingly independent sub-principle to R1 – the data licensing issues are often regarded as out of scope in researcher-initated SKG4RDM implementations, while both the R1 principle and the next two sub-principles (R1.2 and R1.3) are regarded as closely related to research goals and are thus better complied.

R1.2: (Meta)data are associated with detailed provenance The R1.2 principle highlights the importance of associating metadata with detailed provenance, allowing potential users of the data to understand its origins and history. To facilitate reuse, it is essential for researchers to provide a clear narrative of the data's origin, including information on who generated or collected the data, how it has been processed, whether it has been previously published, and if it contains data from other sources that have been transformed or supplemented.

By adhering to the R1.2 principle, researchers enable others to properly cite or acknowledge their work, ensuring appropriate credit for their contributions. Ideally, the provenance information should be presented in a machine-readable format, allowing for seamless integration with automated processes and systems. The partially complying scenarios are sufficiently included in those for R1.

R1.3: (Meta)data meet domain-relevant community standards The R1.3 principle emphasizes the importance of meeting domain-relevant community standards to ensure that metadata and datasets are easy to reuse. Adhering to established community standards enables consistency in data types, organization, sustainable file formats, and documentation. Following these standards fosters better collaboration and data sharing within the research community. Similar to R1.2, while most of the implementations are complying with the R1.2 sub-principle, there are partially complying examples, similar to the scenarios for R1.

When community standards or best practices for data archiving and sharing exist, researchers should comply with them. For example, many research domains have minimal information standards such as MIAME (Minimum Information About a Microarray Experiment) [207] or MIAPE (Minimum Information About a Proteomics Experiment) [208], which FAIR data should meet or exceed. While some community standards may be less formal, the primary objective of FAIRness is to publish metadata in a manner that maximizes its usability within the community. In cases where researchers have valid reasons to deviate from the standard good practice, this should be documented in the metadata. It is crucial to note that the FAIR principles do not address quality issues, as the data's reliability depends on the intended application and the user's perspective.

Discussions

In this subsection, we will summatively examine the advancements and limitations of SKG4RDM implementations, based on the reviewed papers in the above sections. We will also identify the general research gaps and outlook in RDM, especially when guided by the FAIR principles.

Advancements in FAIR principles for guiding SKG4RDM The advent of SKG4RDM implementations has brought forth significant advancements in data use ethics, primarily facilitated through improved data connectivity. These implementations have played a transformative role in promoting responsible data sharing by recognizing and crediting data creators and curators, ensuring data attribution, and fostering a culture of ethical data use. The interlinking nature of large-scale SKGs, such as MAG [62], ORKG [181], and AIDA [27], ensures that data collected for research purposes are not only seamlessly connected, but also managed, shared, and reused in an ethically compliant manner. This advancement in technology aligns closely with the FAIR principles, adding a new ethical dimension to data usage. Notably, SKG4RDM makes data provenance both visible and trackable, introducing a new level of transparency into (meta)data discoverability and usability. Instead only briefly mention the name of the dataset or insert a link as a footnote, data users can easily refer to the data provenance in a graph setting, focus on the information they need, and cite the research data without extra effort of looking up the data. By keeping a comprehensive, immutable record of the history of research data, these implementations allow stakeholders to understand and document the genesis of the data, their various transformations, and who interacted with them, even if in traditionally low-resource fields or regions [32], [209].

The implementation efficiency of SKG4RDM has also seen considerable improvement through streamlined operation processes. The development of these SKGs has enabled the automation of many labor-intensive tasks, such as author name disambiguation and field of research classification, traditionally associated with RDM [194]. As a result, stakeholders can focus more on research processes, community service, or impact measurement, rather than the process of data administration. In particular, semantic technologies and machine learning algorithms employed in SKG4RDM have made data curation, integration, and annotation more efficient [29], [201], [210]. They eliminate redundant RDM steps, streamline metadata creation, and automate linking across different datasets. The use of ontologies and controlled vocabularies allows for consistency in data annotation, which accelerates research data discovery and reuse.

Limitations in FAIR principles for guiding SKG4RDM The FAIR-guided implementation and evaluation of SKG4RDM have demonstrated considerable limitations of research data discoverability and usability. Table 2.4 summarizes the two major limitations, namely ease-security trade-off and knowledge complexity, and provides examples among the 41 reviewed papers.

Table 2.4: Limitations of SKG4RDM with examples and corresponding FAIR sub-principles

Limitations	Example SKG4RDM implementations	Corresponding FAIR sub-principle
Ease-security trade-off reduces user interoperability	[27], [28]	A1.1: The protocol is open, free, and universally implementable
	[29], [30]	A1.2: The protocol allows for an authentication and authorization procedure, where necessary
	[31], [32]	R1.1: (Meta)data are released with a clear and accessible data usage license
Knowledge complexity leads to low reusability	[29], [33]	F2: Data are described in rich metadata
	[34], [35]	I2: (Meta)data use vocabularies that follow FAIR principles
	[36], [37]	R1: (Meta)data are richly described with a plurality of accurate and relevant attributes (with R1.2 and R1.3)

The ease-security trade-off is associated with FAIR's own limitation on the high standard of machine-actionability and less focus on ease for stakeholders. It is essential to strike a balance between the ease of access to data and the necessary security measures. However, the present SKG4RDM systems often favor one aspect over the other, resulting in either data that is hard to access due to stringent authentication and authorization processes or data repositories that are vulnerable to security breaches. The specific problems mainly focus on the sub-principles A1.1, A1.2, and R1.1. First, even when a user has data access, because of the protocol information required for publication is different from what's required for reuse data access protocols (A1.1), usability of the research data is reduced [27], [211]. It is understandable that the RDM effort regarding the data access protocol can be institutional or not directly conducted by the researcher who write about the SKG4RDM implementation. However, when a user needs to apply the whole RDM process, if the protocol is missing, it is often hard to replicate even with full access to data and code. In this sense, although the protocol details sometimes seem to be too technical and not required for publication, researchers and other research data stakeholders should still specify them on an external link or their organizational resource pages. Similar to the problem of getting less coverage in the research process as in A1.1, the sub-principle A1.2, which requires authentication and authorization procedure in the data access protocol, is also not or insufficiently complied with [29], [30]. In such cases, authentication and authorization procedures are either assumed or ignored to reduce complexity in academic research development and presentation, resulting in the lack of accessibility of their RDM implementations. Moreover, proprietary data need a higher level of security [31], [32], which leads to less user-actionability because of the manual access granting. Although these examples are not technically violating R1.1, the actual access to data is likely prohibited from any third party users (not within the organization or having no direct collaboration). As such, it is better to provide comprehensive end-user tools to increase research data usability in an indirect way.

At the same time, knowledge complexity leads to low reusability in SKG4RDM implementations, resulting in limited findability (F2), interoperability (I2), and reusability (R1). In terms of findability, while these systems have made strides in machine-readable metadata, the human element is often neglected. The current SKG4RDM implementations often focus on creating robust machine-actionable applications for data discovery but fall short in designing interfaces and metadata descriptions that are intuitive and user-friendly for human users [29], [33]. This disconnect often impedes the optimal use of data repositories by researchers and other stakeholders who may lack advanced technical knowledge. Similarly, the interoperability of data in some SKG4RDM systems is also questionable. Despite being designed to integrate with other data, the practical applicability of them can be limited due to different real-world scenarios, such as qualitative study requirements or crowdsourcing-based variety in data source [34], [35]. Finally, the reuse promise of data is often hard to fulfill due to the domain-focused community standards, which set a knowledge barrier to cross-domain users. The data may be theoretically reusable, but without proper contextual information and detailed guidance on potential applications [36], [37], users often find it difficult to repurpose the data to their specific needs.

2.2 Research Gaps

Because of the volume, variety of formats, and complexity of connections embedded in the scholarly knowledge linked to research data, it is often hard for researchers and RDM units to organize research data for discovery [60], [212]. Recent RDM guidelines emphasize the importance of discoverability and reusability of research data to promote sharing and transparency of scientific findings [7], [8]. The FAIR principles, while ensuring machine readability and actionability in RDM [20], can lead to a set of general and RDM specific problems, from data use ethics and data openness to implementation efficiency, ease-security trade-off, and knowledge complexity.

Table 2.5: Identified research gaps and corresponding research directions

Identified research gaps	**Context and explanations**	**Source**	**Future research directions of SKG4RDM**
Data use ethics	Recognition of data ownership and respecting consent and privacy (data are findable, but stakeholders are not)	General critiques and extensions (2.1.3)	A SKG4RDM system schema and guidelines properly address data stakeholders and include nodes and edges with high discovery and reuse utilities
Data openness	No access or partial access to data (restrictions on data access)	General critiques and extensions (2.1.3)	Not applicable
Implementation efficiency	Cost, time, and resource	General critiques and extensions (2.1.3)	A system and guidelines ensuring SKG4RDM applications are optimized for both system-end and user-end discovery and reuse easiness
Ease-security trade-off	Full access to data, but low interoperability (users lost context)	Limitations of FAIR-guided SKG4RDM (2.1.4)	Technologies and system guidelines promoting user-centric data search with respect to curators' and users' variations in technical skill and familiarity with data
Knowledge complexity	Full access to data, but low reusability (users have insufficient expertise)	Limitations of FAIR-guided SKG4RDM (2.1.4)	Technologies and system guidelines promoting effective data recommendation and reuse guide generation

Table 2.5 summarizes five main research gaps within the field of RDM and underscore the need for innovative solutions. Machine readability and data metrics do not directly and universally help users determine the reuse potential and relevance of datasets [213]. Context and expertise are two main barriers users face in effectively searching for, evaluating, and deciding whether to reuse research data. The corresponding user-centric, rather than machine-actionable, guidelines and system designs are also essential to enhance usability of research data search and recommednation applications. We further analyze these research gaps

below and provide the corresponding principle and implementation overview in the GAUDS system.

Data Use Ethics. A significant gap in RDM is the recognition of data ownership, and the respecting of consent and support. Even though data might be findable, the stakeholders associated with that data often are not adequately acknowledged. This leads to ethical concerns regarding data usage. To address this, the GAUDS system incorporates a SKG that not only structures data for easy retrieval but also ensures that data stakeholders are visibly and accurately represented within the system. The schema and guidelines of SKG4RDM specifically include nodes and edges that are designed to optimize both the discovery and reuse of data while ensuring high ethical standards in acknowledging data ownership and stakeholder consent.

Data Openness. Restrictions on data access represent another critical gap, where data might be visible but not accessible due to various legal, ethical, or proprietary restrictions. While this specific gap does not directly align with the capabilities of the GAUDS system, it highlights the importance of further expansion in data sharing policies that promote data accessibility to users.

Implementation Efficiency. The cost, time, and resources required to implement and maintain RDM systems can be prohibitive. Efficient implementation is crucial for sustainable operations. The GAUDS system addresses this by ensuring that the SKG4RDM applications are optimized for both system-end and user-end. This involves streamlining processes and enhancing the backend algorithms to reduce processing times and resource consumption, thereby improving the overall efficiency of data discovery and reuse.

Ease-security Trade-off. Full access to data often comes at the cost of interoperability, leading to users losing contextual information. This trade-off between ease of access and security is a significant challenge. The GAUDS system tackles this by employing technologies and system guidelines that promote a user-centric approach to data search. This approach respects the varying levels of technical skill and data familiarity among curators and users, ensuring that the system is accessible yet secure, maintaining the integrity and context of data across different user interactions.

Knowledge Complexity & Reusability. Access to data does not necessarily translate into the ability to effectively reuse that data, particularly when users lack the requisite expertise to interpret complex datasets. The GAUDS system addresses this by integrating technologies and system guidelines that not only facilitate data access but also enhance data reusability. This includes the development of effective data search and recommendation heuristics and algorithms and reuse guides that are tailored to the user's needs and the level of expertise, thereby demystifying complex datasets and making them more usable for a broader audience.

Looking ahead, in the evolving landscape of RDM, the need to supplement the dedicated efforts of data archives and institutional repositories is becoming increasingly imperative. These entities are continuously striving to curate, preserve, and share research data, thereby facilitating its reuse. However, we must make it more convenient for users to explore and acquaint themselves with the available research data. Herein lies the importance of two crucial elements to operationalize user-centric RDM: discovery and reuse.

While a wealth of information is readily available, finding the most relevant datasets for a particular research question can be a challenging task. This is where the SKG-based dataset explorer can be a game-changer. As a novel approach to research data management, discovery, and reuse, SKG can facilitate the efficient discovery and reuse of research data. It can capture, organize, and leverage intricate relationships between data entities, making the exploration and understanding of available research data more accessible. By aligning this power with the user-centric design of research data search and recommendation applications, we can significantly streamline the process of data discovery and reuse, catering to the specific needs of researchers and other data users.

Chapter 3

Conceptualizing Guiding Principles and Prototyping for Data Search Systems

In this chapter, we embark on conceptualizing the foundational principles that guide the development and evaluation of effective data search systems, specifically tailored for the GAUDS system. Stemming from the research gaps outlined in Chapter 2, we introduce a structured set of guiding principles, Connectivity, Effectiveness, Visibility, and Interactivity, collectively referred to as CEVI principles. These principles are crafted to ensure that data search systems not only function optimally but are also deeply aligned with the practical needs of researchers and the broader imperatives of data reuse and management.

The CEVI framework serves as a strategic blueprint that addresses the intricate dynamics of user interaction with data systems, enhancing the ease of access to data and ensuring that these interactions are meaningful and productive. By emphasizing Connectivity, we aim to bridge disparate data sources, enabling seamless access to related information across different datasets. Effectiveness focuses on the accuracy and relevance of search results, ensuring that users retrieve the most relevant data. Visibility improves the transparency of data processes and availability, making it easier for users to understand and navigate the system. Lastly, interactivity ensures that the system is responsive to user inputs and adaptable to varying user needs, facilitating a more engaging and intuitive user experience.

Within the scope of this chapter, we not only discuss these principles in theoretical terms but also demonstrate their practical implementation through the development of two prototypes: SimSearch and DataChat. The SimSearch prototype embodies the Connectivity principle by utilizing advanced semantic similarity techniques. This approach involves measuring the semantic coherence between the metadata of datasets in institutional bibliographies and publications in large-scale scholarly databases. In doing so, SimSearch enhances connectivity within scholarly knowledge, allowing for a more intuitive and autonomous data discovery process that is deeply rooted in natural language understanding.

Following the introduction of SimSearch, we delve into the DataChat prototype. This innovative tool adheres to all the principles outlined in CEVI, with a particular focus on Visibility and Interactivity. DataChat signifies a move toward more dynamic and user-centric data interactions, allowing users to navigate data repositories via a conversational interface. This prototype revolutionizes data search, making it more accessible and tuned to the specific queries and contextual needs of the user, thus greatly improving the overall user experience.

Through these prototypes, the chapter illustrates how the CEVI principles can be practically applied to create more robust, user-centered data search systems. These prototypes not only serve as test beds for the theoretical concepts discussed, but also provide tangible examples of how data systems can evolve to meet the demands of modern research environments. These prototypes also serve to pinpoint practical challenges—such as inflexibility, entity ambiguation, and schema misalignment—that require further refinement in the development of the GAUDS system as a comprehensive, effective, and user-friendly platform.

3.1 CEVI Principles: Connectivity, Effectiveness, Visibility, and Interactivity

To address the research gaps identified in Chapter 2, we introduce a set of novel guiding principles. These principles, collectively referred to as CEVI—Connectivity, Effectiveness, Visibility, and Interactivity—are integral to ensuring that the data search system is not only functional but also user-centric and aligned with the needs of research and data reuse practices. The design and implementation of the GAUDS system is underpinned by the CEVI framework to optimize RDM and enhance user experience. The CEVI principles mitigate the current gaps in data users' expertise and familiarity with data, as well as ensuring the recognition of contributions and supervision needs of data stakeholders, which brings together data providers and consumers in a seamless manner.

Connectivity Connectivity refers to the seamless linking of related datasets, publications, and researchers to illustrate the complex interdependencies inherent in scholarly work. By ensuring that all research entities and their connections are well-captured, the system provides a comprehensive view of the dataset-publication landscape. This connectivity facilitates a holistic understanding of the research ecosystem, making it easier for users to find relevant information and foster collaborative opportunities. It helps bridge the gap between isolated data points, creating a more interconnected and dynamic research environment.

Effectiveness Effectiveness is defined as the ability to deliver accurate, relevant, and timely results that align precisely with the research inquiries of users. The principle of effectiveness ensures that the system outperforms traditional search engines by consolidating fragmented information quickly and accurately. The aim is to reduce the time researchers spend on preliminary searches, providing them with the most relevant and up-to-date data, thus enhancing the overall efficiency of the research process.

Visibility Visibility means ensuring that all key elements and functionalities of the system are easily accessible and understandable to users, regardless of their technical proficiency or the device they are using. By improving the visual and navigational ease of the system, visibility ensures that research entities and their interrelations are easily recognized through advanced frontend visualization tools. Improving visibility helps users quickly identify the resources they need and understand the complex relationships between them, promoting a more intuitive and productive search experience.

Interactivity Interactivity refers to the ability of the platform to support dynamic interactions with data, allowing users to manipulate, explore, and visualize information in various ways. Fostering a high degree of interactivity enhances user engagement and allows for a deeper exploration of datasets. It supports the synthesis and application of knowledge through features that accommodate interactive, customizable, and flexible queries. This enables users to derive valuable insights more effectively, transforming the way data is explored and used in research contexts.

Utility of the CEVI Principles The CEVI principles are integral in addressing current data management gaps in research data management and ensuring the GAUDS system meets high standards for effective, user-centric data search and management practices. These principles should indeed be adopted by researchers, data managers, and institutions. They provide a robust framework for enhancing data search systems, ensuring that they are functional, user-friendly, and aligned with the evolving needs of modern research and data reuse practices. By incorporating CEVI principles, users can achieve significant improvements in the efficiency and effectiveness of their data management systems.

In terms of practical application, the CEVI principles can be implemented in various ways. Connectivity can be achieved by ensuring the seamless linking of related datasets, publications, and researchers within their systems, thus creating an interconnected research environment. This holistic approach facilitates collaboration and knowledge sharing, bridging the gaps between isolated data points. Effectiveness can be enhanced by developing mechanisms that deliver accurate, relevant, and timely results tailored to specific research inquiries, thus increasing the efficiency of data searches and reducing the time researchers spend on preliminary searches. Improving visibility involves designing user interfaces that make all key elements and functionalities easily accessible and understandable, regardless of the technical proficiency of users or the devices they are using. By improving the visual and navigational ease of the system, visibility ensures that research entities and their interrelations are readily recognized, promoting a more intuitive and productive search experience. Interactivity can be fostered by creating platforms that support dynamic interactions with data, enabling users to manipulate, explore, and visualize information in various ways. This enhances user engagement and allows for deeper exploration of datasets, ultimately transforming the way data is explored and used in research contexts.

The CEVI principles apply to several key aspects of data and discovery. Connectivity refers to data integration and linking, ensuring related data points are connected to provide a comprehensive view of the research landscape. Effectiveness relates to search accuracy and efficiency, delivering precise and timely search results to streamline the research process. Visibility concerns user interface and experience, enhancing the visual and navigational aspects of the system to make it more user-friendly. Interactivity pertains to data exploration and engagement, enabling users to interact with data in meaningful ways to derive valuable insights. The adoption and application of the CEVI principles by researchers and data repositories can significantly advance the field of research data management and discovery. These principles help build a more effective, user-centric data search and management system, thereby addressing current gaps and setting a robust foundation for future developments.

3.2 SimSearch Prototype: A Word Similarity-based Data Search System

To study how to implement data search, especially under the Connectivity principle of CEVI, we utilize dictionary-based semantic similarity as an first attempt to autonomize data discovery and propose the Simemantic Search (SimSearch) prototype. The SimSearch prototype uses semantic similarity methods to connect datasets and publications in the ICPSR Bibliography[1]. This prototype aims to enhance the connectivity in scholarly knowledge purely through natural language representations. In particular, we measure the similarities of topical semantics between the metadata of "datasets" in institutional bibliographies, like the ICPSR Bibliography, and "publications" in large-scale scholarly databases, like the Dimensions scholarly database[214][2].

Semantic similarity (or topical relevance) is used to measure the semantic coherence between groups of words and expressions, and is usually used in information retrieval and other NLP tasks to filter and ranking query text, and can be broadly use to improve text and image understanding, knowledge organization and retrieval, reasoning, argumentation, and thinking [215], [216]. Based on the definition and usage of semantic similarity, we define **topical semantics** as the discourse homogeneity represented by the overall (average) semantic similarity between representative relations, e.g. between keywords or summary texts.

Although topical semantics is not yet formally defined in previous work, it has been widely used in computational and social research. [217] depends on topic semantics as a second stage to improve information retrieval systems, in which the relevant documents are ordered as topical sets to provide maximum informativeness to the information requestor (data user). [218] proposes to use topic ontologies and semantic similarity data to evaluate information retrieval methods. In addition to the application in information retrieval research, social media studies also use topical semantics. [219] uses topical semantics, through the follow and retweet relationships on a social media platforms, Twitter, to rank users' quality and topical relevance. Similarly, in other sub-fields of social media study, [220] uses topical aspects from microblogging sites to detect tends and analyze analyze public opinions, [221] empirically characterizes the topical specificity of online community forums to assist community management, and [222] uses topical network analysis to study online activities.

In addition, topical semantics related research is a multiverse of text mining, ranging from simple word frequency to word embedding, that have rich technology method advancements for potential applications. [223] introduce topical n-grams as a text mining model that discovers topical phrases. [224] emphasize the importance of the exclusivity of words to topics in text analysis for communicating content. [225] propose topical word embedding models to represent words using contextual word embeddings, as a more expressive representation than (single) word embedding models. These topical semantics-related methods are helpful reference for our development of computational digital curation methods for data bibliography.

[1] We will introduce the ICPSR Bibliography in details in Chapter 4

[2] The Dimensions API provides the description of the range, the selection, and the format of its concept scores, while the generation or the controlled vocabulary that the concepts are based on are not clear. For more information, please visit https://docs.dimensions.ai/dsl/language.html#concepts-relevance-scores

3.2.1 Data and Methods

DAtaset-Publication Semantic Similarity (DAPSS) Algorithm

To develop an automated method for discovering connections between data, the “datasets”, and scholarly products, the “publications”, we create the DAtaset-Publication Semantic Similarity (DAPSS) algorithm. The DAPSS algorithm includes the following three hyper-parameters:

- A Semantic Similarity Function (F);
- A threshold of semantic similarity between terms (Minimum Similarity Threshold min_sim); and
- A threshold of itemset similarity between sets of terms (Minimum Confidence Threshold min_conf).

The DAPSS algorithm captures the connection and quantifies the relevance between a “dataset” and a “publication” based on the topical semantics. In particular, this algorithm calculate the semantic similarity between the topic spaces of their metadata, “subject terms” and “concepts” respectively, to estimate the likelihood that a “publication” may cite a “dataset”. In this section, we first introduce the nuances of finding the hyperparameters that optimize the the DAPSS algorithm. We then evaluate the performance of it based on an example corpus of datasets and publications.

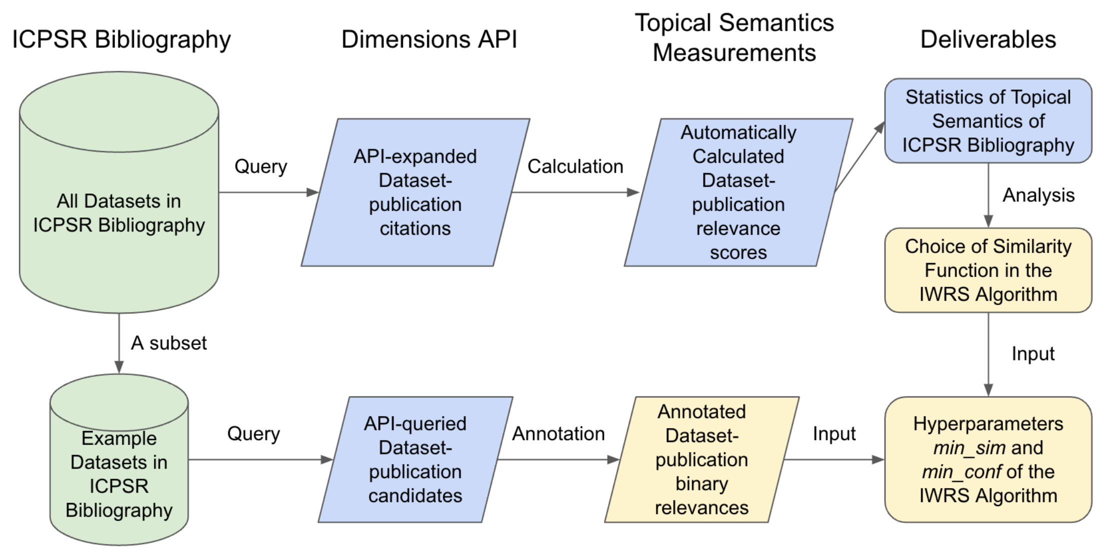

Figure 3.1: The workflow of data processing, algorithms, and analysis

Note: In this figure, we use green color shapes to indicate Bibliography resources, blue color shapes to indicate Dimensions API-scraped resources, and yellow color shapes to indicate data results.

Input: metadata of datasets and publications

The “datasets” are in ICPSR’s Bibliography, which provides a large corpus connecting data and publications that cite them. ICPSR maintains a data archive of more than 250,000 files of research in the social and behavioral sciences, including collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields [226]. The ICPSR Bibliography is suitable for this analysis as it represents a cross-disciplinary collection of several hundred thousand manually-verified data citations.

In specific, except for STUD_NUMS, which is the unique identifier of each Study, we use curated TERMS that contain the alphabetically listed social science subject terms manually associated with each study by ICPSR staff following the Management of Monolingual Thesauri, Z39.19-1993 (NISO 1993). This standard was designed for the multidisciplinary scope and the subject range of the ICPSR archive, including political science, sociology, history, economics, education, criminal justice, gerontology, demography, public health, law, and international relations [227].

The ”publications” are queried from the Dimensions scholarly database that provides a full text search index of over 69 million journal articles and other scholarly works and indexes based on concepts, which further provides the semantic space for matching with the ICPSR Bibliography’s metadata. Dimensions uses machine-learning approach generates, standardizes and reproduces subject categorization of Concepts, enabling researchers to highlight or exclude query terms and to re-rank search results [214].

We study whether we can effectively leverage ICPSR’s manually-created metadata (specifically subject terms attached to datasets) to identify candidate publications within the Dimensions database. For example, Figure 3.2 shows a pair of example dataset-publication with the metadata “Subject Terms” and “Concpets” respectively. The dataset is ICPSR dataset “A comparison of formal and informal dispute resolution in medical malpractice” with “Subject Terms” including ‘arbitration’, ‘conflict resolution’, ‘mediation’, ‘medical malpractice’, and ‘negotiation’ [228]. The publication *A comparison of formal and informal dispute resolution in medical malpractice* cites this dataset and is with metadata “Concepts” including ‘medical malpractice’, ‘malpractice’, ‘comparison’, ‘resolution’, ‘Formal’, ‘dispute resolution’, and ‘informal dispute resolution’. We then calculate relevance scores between a dataset’s subjects terms and a publication’s concepts.

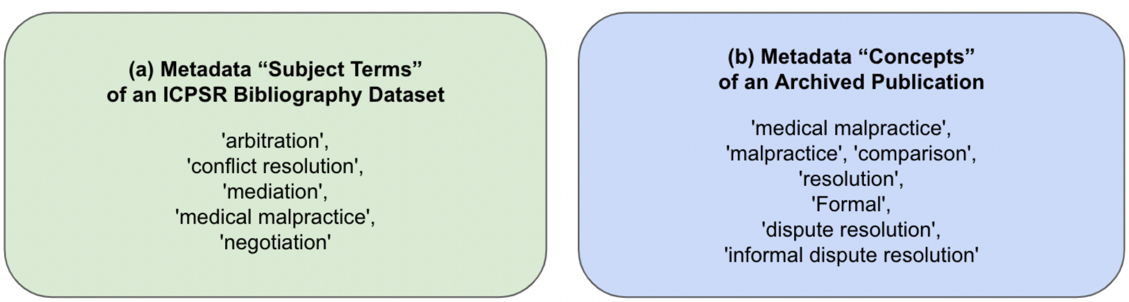

Figure 3.2: Example metadata “Subject Terms” and “Concepts” in a dataset-publication pair
Note: In this figure, we use green color shapes to indicate Bibliography resources and blue color shapes to indicate archived Dimensions API-scraped resources.

Hyperparameter: Semantic similarity measurement function F

Our study uses WordNet and its structured knowledge in word sense disambiguation. WordNet is a widely-used and large-scale English lexical database, which is often used to identify sets of synonyms that supports the task of calculating semantic similarity [229], [230]. Because our study focuses on measuring the similarity, the structure and relations of words in WordNet serve as the fundamental knowledge infrastructure: the structure in WordNet is created by **synonyms**, the words that “denote the same concept and are interchangeable in many contexts”; the structure is then represented as the grouped and unordered sets, the WordNet synsets, which is used to form the **hyperonym** (also known as homonym, super-subordinate relation, and ‘is-a’ or ISA taxonomy or relation); the hyperonym encodes relation among synsets and links more general synsets to construct a hierarchy (also known as a tree), which, for example, for any noun hierarchy, can go up the root node “{entity}” [231].

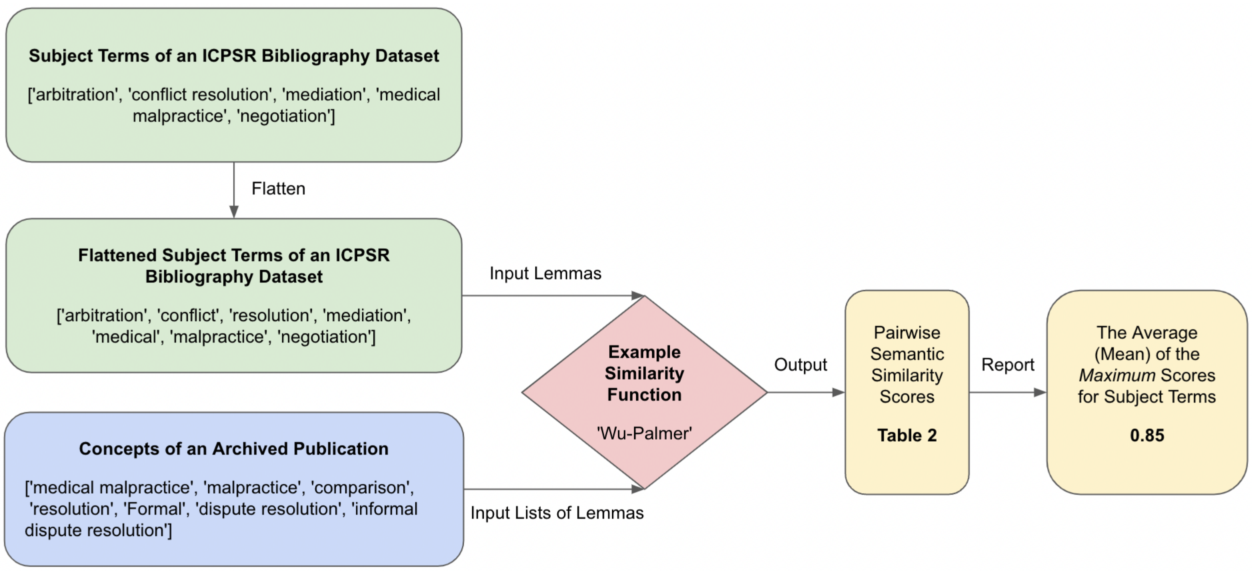

Figure 3.3: The workflow of measuring semantic similarity of the ICPSR Bibliography with an example

Note: In this figure, we use green color shapes to indicate Bibliography resources, blue color shapes to indicate Dimensions API-scraped resources, yellow color shapes to indicate data results, and red color shapes to indicate key algorithms.

Table 3.1: Example pairwise semantic similarity scores

Subject Terms \ Concepts	medical malpractice	malpractice	comparison	resolution	Formal	dispute resolution	informal dispute resolution
arbitration	**1**	0.5	0.9	0.67	0.5	0.56	0.56
conflict	0.71	0.2	0.71	**0.8**	0.5	0.67	0.67
resolution	0.53	0.18	0.53	**0.59**	0.47	0.59	0.59
mediation	0.67	0.18	0.67	**1**	0.47	0.62	0.62
medical	0.59	0.33	0.67	0.67	0.53	**1**	0.67
malpractice	0.53	0.18	0.53	**0.59**	0.47	0.59	0.59
negotiation	0.56	0.25	0.56	0.62	0.47	0.62	**1**

Note: The maximum score of each row is in **bold**.

Using structure and relations provided by the WordNet synonyms and hyperonym, previous study and projects create methods and applications for measuring word sense similarities. Some of them are based on path lengths between synonyms in the hyperonym network, including *path*, *lch* (Leacock-Chodorow) [232], and *wup* (Wu-Palmer) [233]; some others are based on Information Content (IC), including *res* (Resnik) [234], *jcn* (Jiang-Conrath) [235], and *lin* [236]. Although the similarity measuring algorithms are well-developed for word sense similarity, no previous work in using WordNet to measure dataset-publication similarity. In this sense, it is helpful to learn from studies that measure word sense similarity for other units of text, which can be great examples of measuring ranked lists of terms in our case. In general, there are two main approaches: knowledge-based similarity, which represent meanings using "wide-coverage lexical-semantic knowledge resources", and distributional similarity, which "draws on vector space semantics and exploits the statistical distribution of words within unstructured text" [237]. With word sense as the smallest units, previous work also address the similarity measuring issues for short-texts or sentences [238], [239].

Our algorithm focuses on mining the semantic similarity between short-to-mid length concepts and subject terms, which is item-set oriented. Figure 3.3 shows the workflow of measuring semantic similarity of the ICPSR Bibliography with an example. We first prepare the flattened lemmas of subject terms and the concepts of an archived corresponding publication. Taking both lists of word lemmas as the inputs, we then use a similarity function to measure the pairwise semantic similarity scores. As Table 3.1 indicates, each flattened subject term lemma corresponds to a lemmatized concept phrase. We take the best score of all matches in the concept phrase and output it into the cell. Finally, we report the average of the maximum score in each row, as one component (the score for a dataset-publication pair) of the overall topical semantics. To choose the best similarity measurement algorithm, we perform sensitivity analysis on the empirical metadata from the ICPSR Bibliography.

To provide more background information, we summarize the mechanisms of the six algorithms as follows. For the first three algorithms, if we compare a non-empty synset A to another non-empty synset B, we first use the Dijkstra's algorithm [240] to find the shortest path P_{AB} for each ancestor synset common to both synsets; then we measure the similarity of A and B using:

- *path*, the Path Similarity Algorithm calculate the semantic relatedness of two word senses by "counting the number of nodes along the shortest path between the senses in the 'is-a' hierarchies of WordNet" [241]. We can demote the algorithm as:

$$path(A, B) = \frac{1}{P_{AB} + 1} \tag{3.1}$$

- *lch* [232], the Leacock-Chodorow Similarity Algorithm, on top of *path*, incorporates D_A, the depth of the synset A. We can then denote the algorithm as:

$$lch(A, B) = -log\frac{P_{AB}}{2 \times D_A} \tag{3.2}$$

- *wup* [233], the Wu-Palmer Similarity Algorithm, on top of the two previous algorithms, further incorporates D_A and D_B, the depths of synsets A and B; we denote their Least Common Subsumer (LCS) as C, the most specific synset which is an ancestor of both A and B, and the depth of C as D_C. We can then demote the algorithm as:

$$wup(A, B) = \frac{2 \times D_C}{D_A + D_B} \tag{3.3}$$

The rest three algorithms are based on Information Content (IC), which use specific corpora to measure the specificity of a synonymy. The IC resources are derived from the sense–tagged corpora are implemented into the NLTK package – the available IC resources choices including SemCor, Brown, Penn Treebank, British National, while researcher can create their own IC resources [236], [242]. To extract the numerical value that represents the IC for a synset, if we have the IC resource ic and input a WordNet synset Q, then the numerical number extractor function of IC is:

$$IC(Q, ic) = -logP(Q|ic) \tag{3.4}$$

where, for simplicity, we denotes the function as $IC(Q)$. Then, for these three algorithms, to compare two non-empty synsets A and B, which have their LCS C, we can measure their similarity using:

- *res* [234], the Resnik Similarity Algorithm:

$$res(A, B) = IC(C) \tag{3.5}$$

- *jcn* [235], the Jiang-Conrath Similarity Algorithm:

$$jcn(A, B) = \frac{1}{IC(A) + IC(B) - 2 \times IC(C)} \tag{3.6}$$

- *lin* [243], the Lin Similarity Algorithm:

$$lin(A, B) = 2 \times \frac{IC(C)}{IC(A) + IC(B)} \tag{3.7}$$

For the above algorithms, the NLTK package provides WordNet contextualization and implementation specifications [244], [245].

More hyperparameters: similarity thresholds min_sim and min_conf

In addition to and based on the benchmarks set by topical semantic measurements, we create a dataset-publication semantic similarity prediction method that improves curation through automation. In particular, we develop an **Itemset-inspired and WordNet-based relevance scoring (DAPSS) algorithm** that serves as the the core algorithm of phrase-level similarity calculation. Inspired by frequent item-sets calculation algorithms, using a Similarity Measuring Function F, DAPSS algorithm compares WordNet word sense synsets and show their similarity – it measures the relevance between two phrases, and aggregates such measurement of pairs to obtain an overall Relevance Score r, based on the Minimum Similarity Threshold min_sim. The algorithmic details of the DAPSS Algorithm is shown in pseudo code below:

Require: Term Lemmas T, Concept Lemmas C, WordNet Similarity Function F, Minimum Similarity Threshold min_sim

Ensure: Number of Relevant Terms $r = 0$

$M \leftarrow dim(T)$

$N \leftarrow dim(C)$
$S \leftarrow zeros(M)$
for t in $1 : T$, m in $1 : M$ **do**
  $V \leftarrow zeros(N)$
  for c in C, n in $1 : N$ **do**
    $V_n = F(T_m, C_n)$
    $S_m = median(V)$
    if $S_m \geq min\_sim$ **then**
      $r+ = 1$
    end if
  end for
end for
return r

As Figure 3.1 shows, the Concept Lemmas C are the list of word lemmas of Dimension API returned concepts in from more relevant to less relevant order (with their comparative relevance score), and the terms are lists of curated word lemmas from the ICPSR metadata.

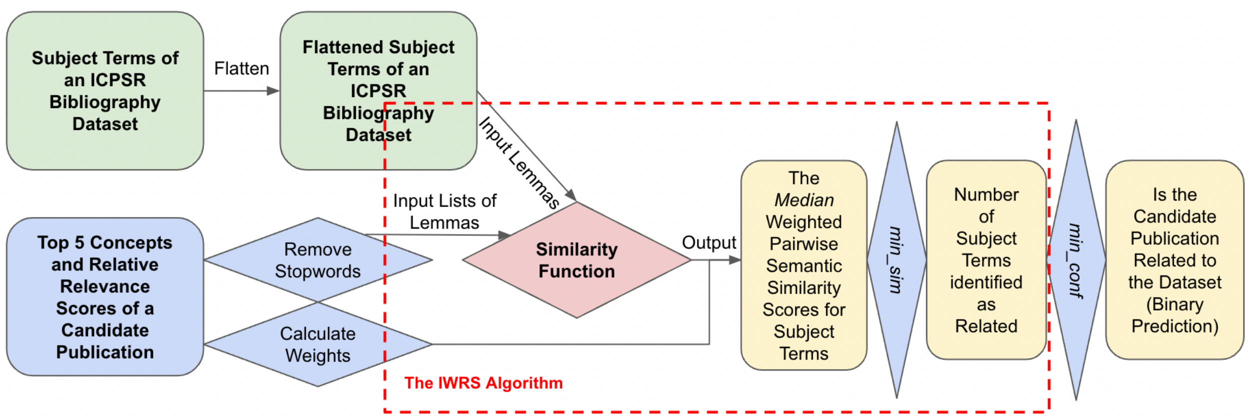

Figure 3.4: The workflow of the extended DAPSS algorithm

Note: In this figure, we use green color shapes to indicate Bibliography resources, blue color shapes to indicate Dimensions API-scraped resources, yellow color shapes to indicate data results, and red color shapes to indicate key algorithms.

The DAPSS algorithm successfully serves as a detector with scores the the outputs, but it cannot provide binary suggestions as a Filter or Detector as the Funnel Process required. As Figure 3.4 shows, there are two values that need to be predefined to make the additional functionalities work. For each similarity measuring functions F, we require hyper-parameters the Minimum Similarity Threshold (min_sim) and the Minimum Confidence (min_conf), where min_conf provides the percentile threshold to define if a publication is relevant to a dataset.

To study these hyper-parameters, we annotate a dataset to help the model development of the DAPSS algorithm. We manually inspect 500 queried papers from the Dimensions API as data citation candidates for 12 ICPSR "data" (as shown in Figure 3.5), where we make binary annotations and result in an annotated dataset with 362 valid samples[3]. Based on the full metadata of those "data" and "publications", as well as the full text of the publications, there are 182 samples annotated as "relevant" (with code 1) and 180 samples annotated as "irrelevant" (with code 0).

	STUD_NUMS	ALL_NAMES	DOI	icpsr_PUBS	STUDY	NAME	TERMS	DESCRIPTION
0	7010	[ANES 1972 Time Series Study, American Nationa...	10.3886/ICPSR07010	1239	7010	ANES 1972 Time Series Study	candidates; congressional elections; cultural ...	This study is part of a time-series collection...
1	34315	[The Irish Longitudinal Study on Ageing (TILDA...	10.3886/ICPSR34315	392	34315	The Irish Longitudinal Study on Ageing (TILDA)...	activities of daily living; aging; depression ...	The Irish Longitudinal Study on Ageing (TILDA)...
2	36692	[Chinese Longitudinal Healthy Longevity Survey...	10.3886/ICPSR36692	270	36692	Chinese Longitudinal Healthy Longevity Survey ...	activities of daily living; aging; alcohol con...	The Chinese Longitudinal Healthy Longevity Sur...
3	36231	[Population Assessment of Tobacco and Health (...	10.3886/ICPSR36231	180	36231	Population Assessment of Tobacco and Health (P...	adults; advertising; alcohol; bidis; chewing t...	The PATH Study was launched in 2011 to inform ...
4	3818	[Harvard School of Public Health College Alcoh...	10.3886/ICPSR03818	86	3818	Harvard School of Public Health College Alcoho...	academic achievement; alcohol abuse; alcohol c...	This survey interviewed students at colleges t...
5	3163	[Harvard School of Public Health College Alcoh...	10.3886/ICPSR03163	79	3163	Harvard School of Public Health College Alcoho...	academic achievement; alcohol abuse; alcohol c...	This study resurveyed colleges that participat...
6	4291	[Harvard School of Public Health College Alcoh...	10.3886/ICPSR04291	71	4291	Harvard School of Public Health College Alcoho...	academic achievement; alcohol abuse; alcohol c...	The Harvard School of Public Health College Al...
7	6577	[Harvard School of Public Health College Alcoh...	10.3886/ICPSR06577	61	6577	Harvard School of Public Health College Alcoho...	academic achievement; alcohol abuse; alcohol c...	This survey focused on alcohol use and alcohol...
8	2559	[Eurobarometer 49: Food Product Safety, Child ...	10.3886/ICPSR02559	24	2559	Eurobarometer 49: Food Product Safety, Child S...	attitudes; cancer; child prostitution; disease...	This round of Eurobarometer surveys queried re...
9	7706	[Juvenile Detention and Correctional Facility ...	10.3886/ICPSR07706	9	7706	Juvenile Detention and Correctional Facility C...	census data; correctional facilities (juvenile...	The 1974 census includes juvenile detention an...
10	20342	[Census of Medical Examiners' and Coroners' Of...	10.3886/ICPSR20342	5	20342	Census of Medical Examiners' and Coroners' Off...	autopsy; coroners; death records; forensic med...	The Census of Medical Examiners' and Coroners'...
11	36530	[Survey of Midlife in Japan (MIDJA 2): Biomark...	10.3886/ICPSR36530	2	36530	Survey of Midlife in Japan (MIDJA 2): Biomarke...	anger; anxiety; biomarkers; chronic illnesses;...	In 2008, with funding from the National Instit...

Figure 3.5: Examples of "data" and "metadata" in the ICPSR curated studies

To tune min_sim and min_conf for the DAPSS algorithm, we use the grid search method that goes over the proper ranges, where the goal is to find the best combination min_sim and min_conf values that lead

[3]By saying "valid", we refer to the English language publications that have non-empty extracted concepts, as well as accessible links to their full texts (in PDF format).

to the highest balanced score [246]. We use the confusion matrix and scores derived from it to define the balanced accuracy score as

$$\frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right) \tag{3.8}$$

where P is the Positive prediction, N is the Negative prediction, TP is True Positive prediction, and TN is True Negative prediction. We use the implementation of the balanced accuracy score function ($balanced\_accuracy\_score$) in Scikit-learn [247] to tune these hyper-parameters.

3.2.2 Evaluation: Simplicity and Flexibility of the DAPSS Algorithm

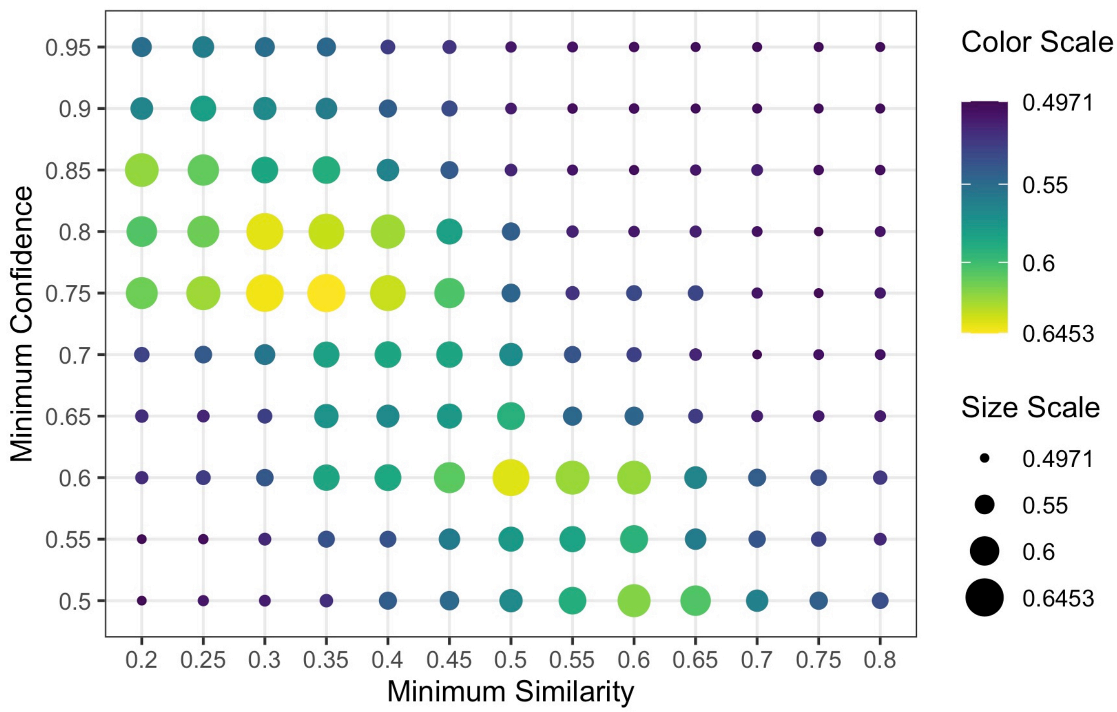

Figure 3.6: Model tuning results on grid search of balanced accuracy scores

Figure 3.6 shows the grid search of the two hyper-parameters of the extended DAPSS algorithm that predicts the binary classification of data-publication relevance. The grid, with Minimum Similarity (min_sim) as the horizontal axis, Minimum Confidence (min_conf) as the vertical axis, and Balanced Accuracy Score as the size and color of the nodes, considers all hyper-parameter combinations of the sense-making range. For min_sim, the grid starts at 0.2 and ends at 0.8 to avoid extremely similar or different matching scores of WordNet synsets. For min_conf, the grid starts at 0.5 and ends at 0.95 because the confidence is too low if less than a half of the itemsets are matching. Through the iteration of a 10×13 grid, we find the combination $min\_sim = 0.35$ and $min\_conf = 0.75$ yields the best balanced accuracy score, 0.6453, for the Resnik algorithm.

The DAPSS Algorithm and its expansion supports the discovery of informal data citations in publications through measuring topical semantics (Figure 3.3) and predicting data-publication relevance (Figure 3.4).

Zooming in to the DAPSS algorithm helps us understand the simplicity and applicability of the prediction method. As previously introduced, there are three main hyper-parameters of the extended DAPSS algorithm, including a WordNet-based similarity measurement function F and the thresholds of the minimum term-level similarity min_sim and minimum itemset-level confidence min_conf. For the choice of the semantic similarity function F, we did a thorough comparison between Resnik and Leacock-Chodorow algorithms (in Section 3.2.2). The fields of digital curation and science of science research are not traditionally computation-intensive, which result in less than adequate resources and comparatively expensive computing power. Thus, we choose the Resnik algorithm because the two algorithms are almost equally well-performing in accuracy and variance (Table 3.2), while the Resnik algorithm has a much lower processing time requirement. For the choice of the two hyper-parameters of thresholds, we refer to the Grid Search tuning results, where $min\_sim = 0.35$ and $min\_conf = 0.75$ yield the best balanced accuracy score (the most color-wise intensive yellow node in Figure 3.6). We should also notice that there are several similar color-wise yellow-green nodes, like $min\_sim = 0.3$ and $min\_conf = 0.8$ (close the best node) or $min\_sim = 0.5$ and $min\_conf = 0.6$ (distant from the best node). These additional good performance nodes confirms the legitimacy of the overall design of the DAPSS algorithm and its extension – they mostly locate in a negatively associated line in the grid, which indicates the non-randomness of the algorithmic design. In addition, the close but not the best node ($min\_sim = 0.5$ and $min\_conf = 0.6$) shows the complexity of word sense matching, as the balanced accuracy score changes for $min\_conf = 0.6$ is not linear, which also confirms the intuition of searching a wider range of the minimum similarity threshold.

While there are overlapping steps in both workflows, their individual goals corresponds to different digital curation needs. The topical semantics measurement workflow stops at reporting the average maximum scores, while the prediction workflow focuses on predicting individual data-publication relevance, which requires further steps and thresholds. Moreover, we use different statistics measurements in reporting topical similarities. The topical semantics measurement workflow uses the *maximum* scores to represent each row of the pairwise semantic similarity score matrix, which is based on the fact that the ICPSR Bibliography contains high-quality curated metadata. The DAPSS algorithm and its extension then uses *median* scores in for reporting pairwise semantic similarity, as it handles individual data-publication relevance, which requires a better weighted statistic that avoid extreme values. This reporting mechanism avoids tedious pair-wise comparison of long text, which saves computation resources and can be widely adopted by smaller research or curation teams.

Beyond the internal design of the DAPSS algorithm, the input data also enables the unique advantages of this text mining method. First, Funnel Process (Figure 3.1) enabled by the DAPSS algorithm is positioned after the step of Import Citations and before storing to database in the current ICPSR data citation curation process, which brings as much prior domain knowledge into the algorithm as possible to ensure its prediction quality. At the same time, it avoids the intuitive relevance judgement as a potential problem and maintains the possibility of finding “surprising” or novel instances of data reuse. In addition, this method helps avoid the access limitations of publications’ full text. The use of metadata for coherence analysis, instead of matching entities from full-text, expands the scope of bibliography curation – candidates suggested by the Funnel Process may not have publicly available full-text, but can then be manually found by bibliography curators, if suggested as a curation priority. In this sense, the Funnel Process optimize the Bibliography curation process by having human and computer efficiently collaborate.

3.2.3 Case Study: Using DAPSS Algorithm to Evaluate the ICPSR Bibliography

We use semantic similarity functions in the DAPSS algorithm as metrics to evaluate the digital curation practice in the the ICPSR Bibliography. We study the metadata of 101,144 ICPSR Bibliography that have a DOI (about 85% of the whole bibliography) and then analyzes the semantic similarity between Dimensions API's concepts scores and ICSPR subject terms. We use average similarities (median and mean), variances (overall and by dataset, and processing time (runtime in minutes), to conduct a sensitivity analysis. The purpose of the sensitivity analysis as to find out the best similarity measurement function as well as the corresponding average similarity score, as both the topical representation of the ICPSR Bibliography and a threshold for choose the semantic similarity measuring method.

We measure topical coherence of the ICPSR Bibliography through the six similarity measurement algorithms, including the Path Similarity algorithm, the Leacock-Chodorow Similarity algorithm, the Wu-Palmer Similarity algorithm, the Resnik Similarity algorithm, the Jiang-Conrath Similarity algorithm, and the Lin Similarity algorithm, where the later three algorithms functioning with Information Content generated from the "genesis" NLTK corpus. We use each of these six similarity measurement algorithms to calculate the topical semantics of each dataset, i.e. the semantic similarities between each dataset's subject terms and each of its citing publication's concepts.

Table 3.2 shows both the measurement and the sensitivity analysis results. First, the Overall Median of 0.658 and Overall Mean of 0.65 show the high-level thematic matching between ICPSR studies and archived publications. The Resnik and Leacock-Chodorow algorithms both support more than 90 percent topical coherence, which shows high level the solidarity in curatorial actions. Moreover, the Overall Variance, Variance of Variance by Dataset, and the Processing Time together serve as an empirical sensitivity analysis that compares the performance of the six algorithms. Regarding the performance variances on accuracy, the Leacock-Chodorow algorithm has only 0.0059 overall variance and 0.00012 variance by dataset, which is the most consistent algorithm. However, by taking into consideration of Processing Time, where the Resnik algorithm is empirically proved to be the most efficient, the latter algorithm does a overall better job in balancing variance and accuracy. Since the Resnik algorithm is only 0.0004 and 0.00002 less stable in overall and dataset-wise accuracy, while it process more than 4 times faster than the Leacock-Chodorow algorithm, we finally choose to use the Resnik algorithm for further steps and overall topical semantics representation.

To evaluate the overall coherence of the archived data-citing papers, we compare the overall accuracy of the topical semantics of the ICPSR Bibliography, which is more than 0.65, to the baseline accuracies of knowledge-based semantic similarity methods. Since our topical semantic measurement methods is based on WordNet synsets, a knowledge-based word sense corpus, it is better not to use negative sampling methods to draw random samples from the current ICPSR Bibliography. Instead, we compare the performances of the current method to three related empirical assessments of topical similarity measurements. Using knowledge-based semantic similarity methods and supervised learning approach, [215] empirically reaches the overall accuracies of 0.715 for the paraphrase identification task and 0.586 for the entailment identification task.

Table 3.2: Measurement results and sensitivity analysis of the six similarity algorithms

Similarity Measurements	*path*	*lch*	*wup*	*res*	*jcn*	*lin*	**Overall**
Overall Median	42.76%	90.00%	77.1%	**90.90%**	38.00%	56.09%	65.81%
Overall Mean	43.32%	89.60%	76.01%	**89.97%**	38.28%	55.19%	65.06%
Overall Variance	0.0156	**0.0059**	0.0093	0.0063	0.0236	0.0206	0.01355
Variance of Variance (by Dataset)	0.00025	**0.00012**	**0.00012**	0.00014	0.00055	0.00045	0.00027
Processing Time (in Minutes)	1911	1149	2148	**225**	234	227	982.33

Note: The best performance in each metric is labeled in **bold**.

[223] uses topical bag-of-words models to for an information retrieval task and result in average precisions of 0.1996 for the the bigram topic model, 0.2107 for the LDA collocation model, and 0.2122 for the topical n-gram model. [224] compares the classification performance of the Hierarchical Poisson Convolution (HPC) model, which ranges from 0.332 to 0.711 based on different sub-task and base-model combinations. Although we cannot use these previous accuracies as the strict baseline, as our ICPSR Bibliography is unique and our tasks are different, we do can refer to their overall performance to evaluate the solidarity of topical semantics in the ICPSR Bibliography.

High topical coherence implies effective criteria and complete curation. The above 0.65 overall accuracy in the benchmarks shows a high level of coherence between the data-citing publications and the datasets in the ICPSR Bibliography. The overall assessment of topical semantics not only benefit the DAPSS algorithm as a method in the later research question, but also other ICPSR datasets with low curation rate or fewer number of data citations. The overall coherence of the archived data-citing papers provides soft thresholds for newly queried candidate publications. Especially for the less-curated datasets with automatically curated data-citing publication candidates, this threshold can be used to suggest curatorial priorities and to provide high-level understanding of the bibliography. In addition, the coherence scores can be used to characterize and compare the diversity of reuse among the collected citations. To summarize, the high overall topical coherence score here implies strict curation criteria of the ICPSR Bibliography and the accomplishment of high quality curatorial actions.

3.2.4 Discussion

Topical semantic analysis and DAPSS algorithm are related but different semantic similarity methods. While the overall topical semantics of the ICPSR data is proved to be cohesive, we observe different empirical prediction performance when it comes to unseen data and metadata. Because of the unequal curation resources for different datasets, the DAPSS algorithm and its extension depend on additional parameters and hyper-parameters to make better predictions. In this section, we direntify both the contribution of DAPSS and the flexibility challenge.

DAPSS's contribution to the general problem of data search

Evaluating and enhancing digital curation of data archives' bibliographies, for example, the ICPSR Bibliography, benefit both curators and researchers. For curators, analysis and automation of curatorial actions can help institution data archives to better allocate curatorial resources; for researchers, the automated or semi-automated process of data citation network detection can reveal information about data use and its importance in producing scientific knowledge, which are emerging topics in science of science and data work research.

DAPSS leverages semantic similarity methods to represent the topical coherence of the curated bibliography of the ICPSR data archive, and develop the DAPSS algorithm and its extension to assist future curatorial work. Through grid search tuning, this algorithm is an unique prototype for similar bibliography curation tasks, which may have limited computation power, low-resource corpora, or hard-to-find full text.

The Flexibility Challenge in the Word Similarity-based Data Search

While word similarity-based data search and data linkage improvements can potentially link data together, it remains challenging for data archives to rely on the current system to enable data discovery and reuse due to its limitation in **flexibility**. This limitation stems primarily from issues of flexibility inherent in WordNet. The system's reliance on empirical experiments to decide on the proper similarity measurement algorithms and thresholds can lead to imprecise or incorrect predictions, particularly in domain-specific contexts.

We provide examples in Figure 3.7 of prediction results from the ICPSR dataset series "Harvard School of Public Health College Alcohol Study" [248] with detailed subject terms shown in the metadata (a). The predictions are categorized into True Positive (Figure 3.7 (b)), True Negative (Figure 3.7 (c)), False Positive (Figure 3.7 (d)), and False Negative (Figure 3.7 (e)). True Positive (TP) predictions successfully identify relevant concepts like 'unhealthy alcohol use' and 'general health assessment' that directly correlate with the dataset's focus on alcohol consumption and health screening. True Negative (TN) predictions correctly exclude unrelated concepts such as 'financial adversity' and 'intimate partner violence,' demonstrating the system's ability to distinguish non-relevant terms despite sharing some context like 'alcohol use.' False Positive (FP) predictions inaccurately associate concepts like 'pre-pregnancy smoking' with the dataset, likely due to superficial word similarity such as 'smoking' appearing in both the dataset's focus and the falsely linked concepts. False Negative (FN) predictions miss relevant connections, such as 'visualization software CiteSpace,' which, while relevant to data analysis, is incorrectly dismissed probably due to its semantic distance from typical health-related terms.

As the examples shown, the current method utilizes resources like WordNet and related corpora as the lexical base, which is a finite and static resource. This limitation results in less coverage of newly created words and their senses, inadequate for measuring similarities in datasets and publications covering new or evolving topics. This could lead to inclusiveness and diversity issues in the algorithm's development, as it may not fully represent terms related to emerging social or scientific domains.

In addition, the effectiveness of the prediction system is compromised in domain-specific scenarios. For example, the inclusion of 'visualization software CiteSpace' as a False Negative prediction highlights the

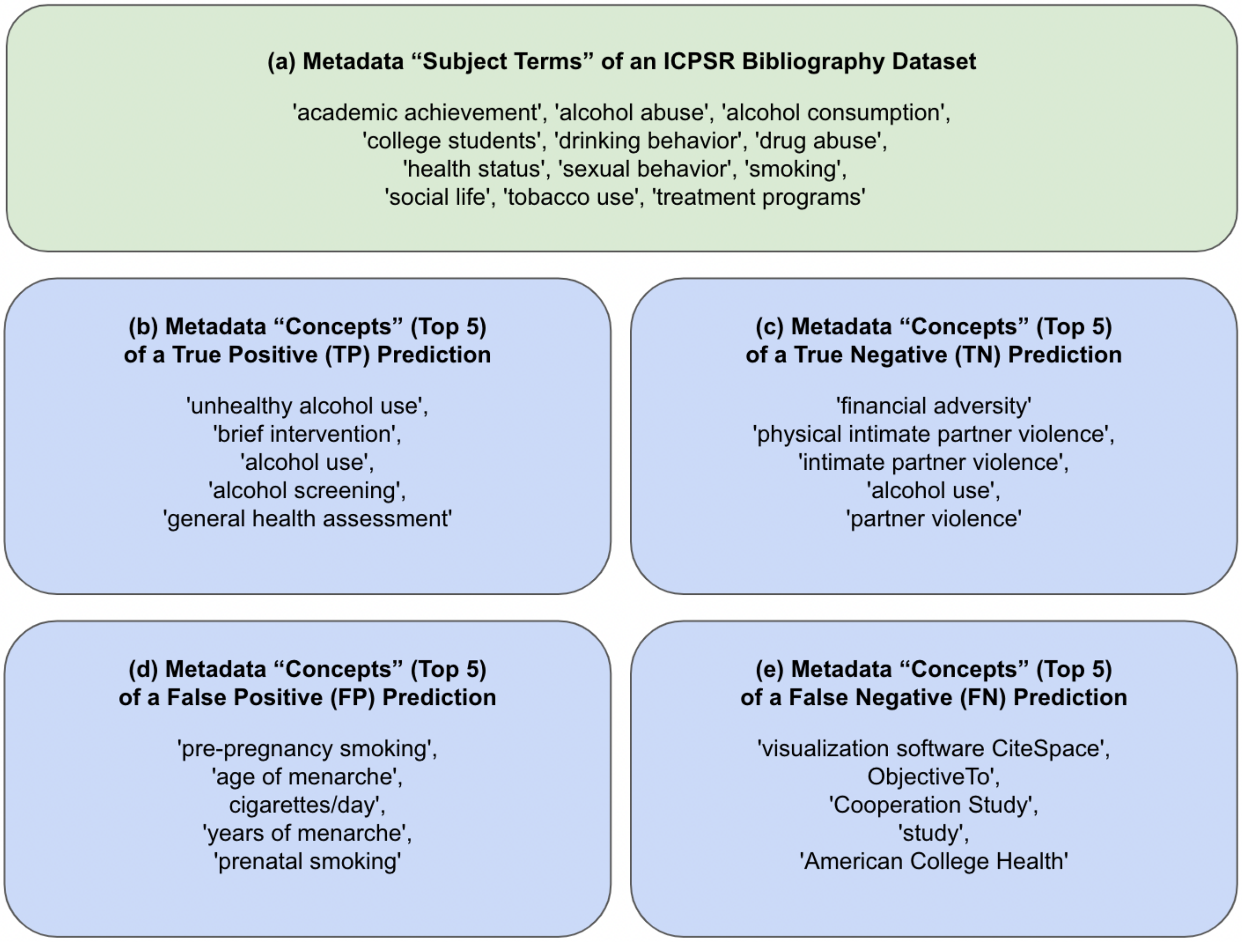

Figure 3.7: Example metadata “Subject Terms” and “Concepts” of dataset-publication relevance prediction

Note: In this figure, we use green color shapes to indicate Bibliography resources and blue color shapes to indicate archived Dimensions API-scraped resources.

system’s shortcomings. It is marked irrelevant despite its potential applicability to the dataset, likely due to a lack of contextual understanding where the term ‘software’ was misaligned with the expected domain-specific terminology.

The illustrated examples underscore the need for enhancing the word similarity-based data search and linkage systems to improve their flexibility for better effectiveness, especially in accommodating newly emerging terms and maintaining contextual relevance in predictions. This advancement is crucial for facilitating more accurate data discovery and reuse in dynamic and diverse research fields.

3.3 DataChat Prototype: A SKG-based and LLM-augmented System for Dataset Search and Visualization

To further study how to build an autonomous and user-centric system for data search, guided by the CEVI principles, we prototype a chat-like data search system called DataChat. To facilitate DataChat, we developed an SKG for ICPSR (ICPSR-SKG) that encodes the same metadata currently available through the ICPSR search system and enables new interactions with research datasets in three main ways. First, the ICPSR-SKG explicitly stores context about the relationships between entity types (e.g., publications and datasets) that users can access, explore, and query. Second, the SKG renders interactive network visualizations, which support user understanding of large-scale relationships across entity types. Finally, unlike systems that are built on static indexes, the SKG is built on top of a graph database, which supports natural language understanding that leverages the connections within the data.

DataChat uses the same underlying metadata currently available in ICPSR's dataset search to contribute novel: (1) front-end interactions for users (i.e., natural language queries and network visualizations); and (2) back-end relationships in databases (i.e., semantic triples). As a conversational assistant to dataset users and other stakeholders, DataChat traverses ICPSR-SKG as the knowledge base for answering users' dataset-related questions. DataChat then presents the resulting textual and visual representations in an interactive user interface, enabling users to explore relationships between research datasets available from ICPSR.

3.3.1 Data and Methods

We selected the DataChat technology stack shown in Figure 3.8 based on our original design goals of: (1) enhancing metadata context by exposing links between entity types; and (2) increasing users' proficiency with the search system, regardless of their level of research expertise. The search system is centered around Datasets, which have explicit contexts derived from their relationships with other scholarly entities, including research Publications.

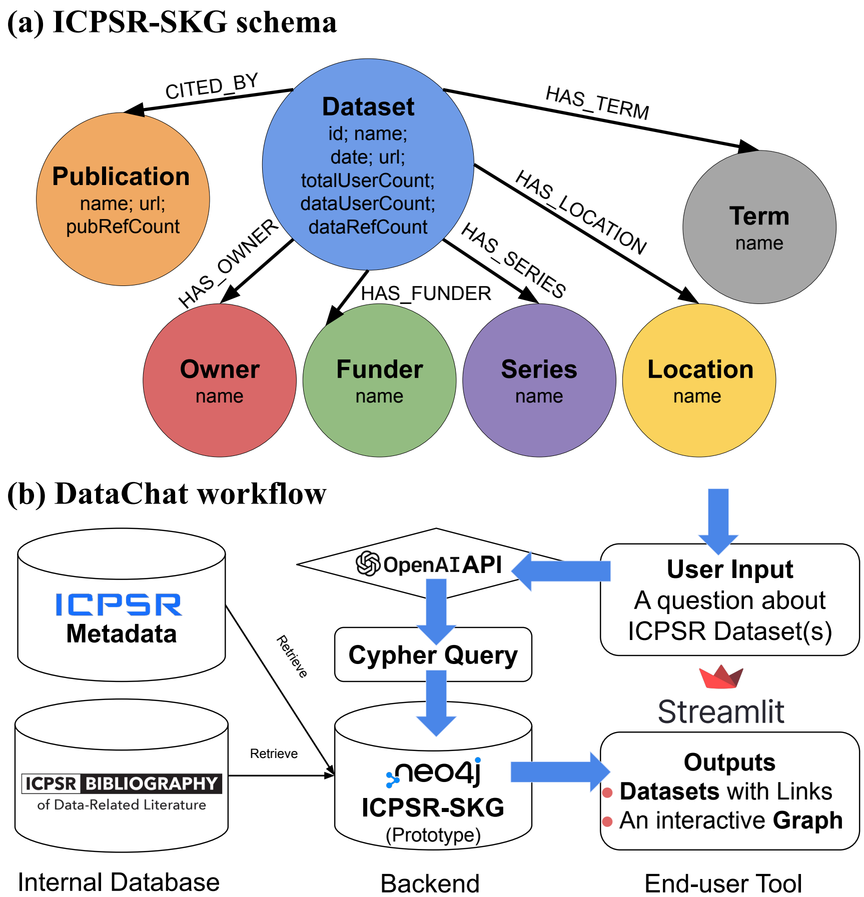

Figure 3.8: DataChat design: (a) schema and (b) workflow for the ICPSR-SKG graph database prototype

As Figure 3.8(a) indicates, the schema of the ICPSR-SKG prototype includes dataset nodes and other types of nodes linked to them. For scalability and experimentation, we selected the 1,642 ICPSR datasets released from 2017 to 2022. Dataset nodes have seven attributes, including the dataset's unique identifier ("id"), its formal study title ("name"), its creation "date", the "url" of its DOI, the total number of users who downloaded any metadata or data of the dataset ("totalUserCount"), the number of users who downloaded datasets ("dataUserCount"), and total number of publications that have cited the dataset ("dataRefCount"). The other six types of nodes are linked to dataset nodes through unique types of relations. While all six types of nodes, including publication, owner, funder, series, location, and term, have the "name" attribute, the publication nodes also have the "url" of DOI and the number of citations ("pubRefCount"). We derived publication information from the ICPSR Bibliography [249].

Figure 3.8(b) illustrates the DataChat system design, incorporating a seamless workflow between an end-user tool based on Streamlit [250], a backend processing system utilizing the OpenAI API [251], and an internal Neo4j-based ICPSR-SKG retrieving data from ICPSR databases [39]. The interaction process starts with the user input, a natural language question about datasets, on the Streamlit interface, which is then sent to the backend for processing using the OpenAI API of the GPT-3.5-turbo model [252]. The API processes the prompt to generate a Cypher query, the native query language for Neo4j databases, where the prompt is based on the combination of the user input and engineered input-output pairs. We provide an example pair of input and ouput as follow.

Corresponding Cypher query output MATCH (a:Dataset) WHERE a.owner <> 'ICPSR' RETURN a.name + " LINK: " + a.url AS response ORDER BY a.dataRefCount DESC LIMIT 5

The ICPSR-SKG Neo4j database then executes the generated Cypher query to retrieve relevant nodes and edges, which are returned to the Streamlit-based interface as either chat messages or a subgraph of the ICPSR-SKG using the streamlit-agraph, a Streamlit Python package that visualizes interactive network graphs [253].

3.3.2 Results

The DataChat dashboard includes two tabs, the DataChatBot Tab and the DataChatViz Tab. Based on users' natural language inputs, these two tabs respectively provide suggestions of datasets with links and visualize interactive graphs for users' exploration.

To evaluate the performance of the DataChat prototype, we generated and tested 105 natural language questions about ICPSR datasets. These questions were inspired by a prior study of "genuine information needs" [254] for specific social science data stakeholder perspectives from education, funding agencies, and data management units. These questions provide a preliminary evaluation of DataChat's overall ability and versatility.

Interface

Figure 3.9 shows the results of the DataChat dashboard for the example input "What are the latest datasets owned by ICPSR that have been cited by publications more than 3 times?" We intentionally used a grammatically ambiguous query to demonstrate the system's flexibility in query interpretation. The DataChatBot Tab (Figure 3.9(a)) contains three parts, including a question input frame on the bottom, the conversation panel on the top left, and the generated Cypher query on the top right. Users can modify the input and rerun by pressing the "Enter" button. The resulting messages start at the bottom and scroll up, similar to texting, promoting familiarity and ease of use, as most users are already accustomed to this layout. We also keep the Cypher query available to users for transparency, debugging, learning, and feedback purposes.

The DataChatViz Tab (Figure 3.9(b)) is a colored graph visualization where colors correspond to object types (Figure 3.8(a)). In addition to visualizing different node types and names, the graph also highlights the attribute nodes shared by at least two datasets, which are positioned at the center of the graph. For example, the American Health Values Survey and the Massachusetts Health Reform Survey are both owned by HMCA, the Health Management Company of America. Notably, the graph(s) are also interactive – users can highlight and place the nodes and edges for their illustration needs[4].

[4]A video demo of the DataChat dashboard is available on `https://youtu.be/y4EaJzV2nA8`

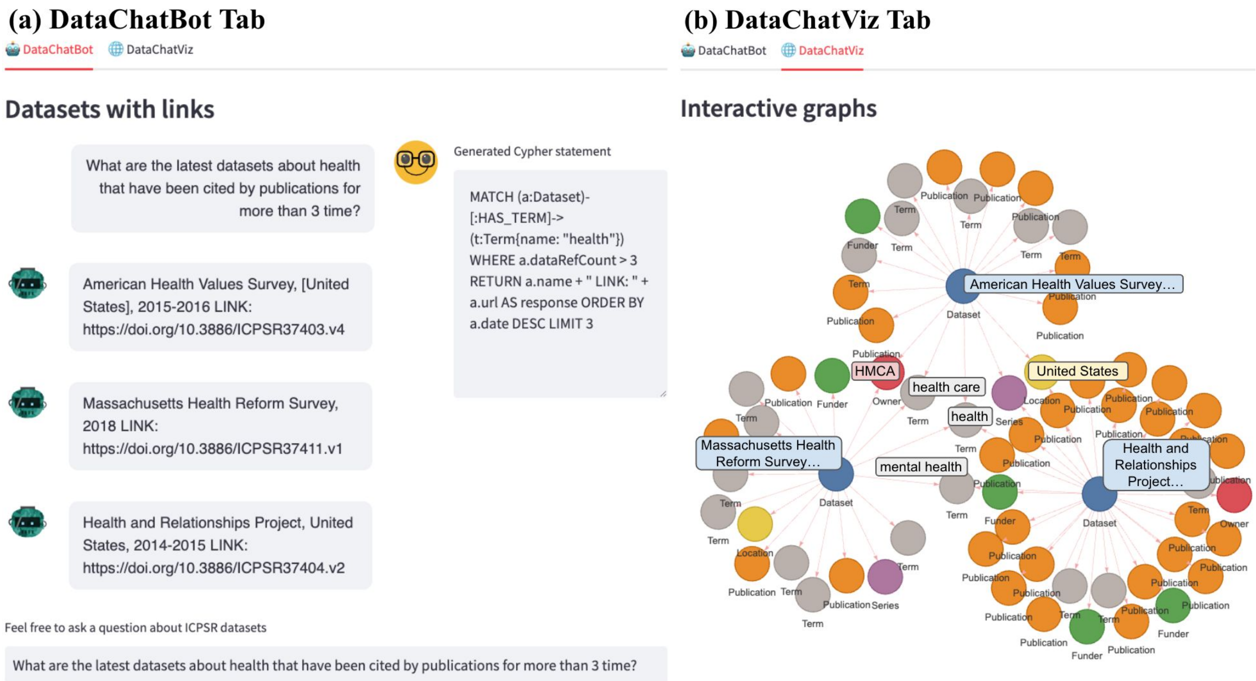

Figure 3.9: Prototype DataChat user interface of ICPSR-SKG

Note: Several mouse-over effects in the interactive tab are combined and presented simultaneously, while the actual graph shows them separately.

Evaluation

Two annotators evaluated the system outputs of the 105 questions and annotated them as "pass" or "not pass". Questions "passed" if they met the expected information need in the original query (semantics of natural language) and compiled a query that can be executed within the existing schema (syntax); they did "not pass" if they returned incomprehensible results or violated the schema. The evaluators reached a high level of agreement in the first round of annotation (α = 0.87) [255], and reached the full agreement after discussion. As Table 3.3 shows, the overall pass rate was 61% (64/105), while the pass rate per stakeholder varied: for education-related (e.g. researchers, educators, and students) queries, the pass rate was 83% which was the highest; for data management-related (e.g. archivists, librarians, and bibliographers) queries, the pass rate was 74%; for funding-related (e.g. governments, NGOs, and corporates) queries, the pass rate was 26%, which was the lowest.

Table 3.3: Evaluation of DataChat with stakeholder-specific examples

Pass rate per stakeholder	Input example	Corresponding Cypher query output
Education: 29/35 (83%)	What are the most popular datasets about mental health?	`MATCH (a:Dataset)-[:HAS_TERM]->(t:Term) WHERE a.name CONTAINS 'mental health' RETURN a.name + " LINK: " + a.url AS response ORDER BY a.dataUserCount DESC LIMIT 3`
Funding agency: 9/35 (26%)	Which datasets have been funded by the National Institutes of Health or Ford Foundation?	`MATCH (a:Dataset)-[:HAS_FUNDER]->(f:Funder) WHERE f.name IN ["National Institutes of Health", "Ford Foundation"] RETURN a.name + " LINK: " + a.url AS response ORDER BY a.date DESC LIMIT 3`
Data management unit: 26/35 (74%)	Which datasets include information from countries in the Middle East, such as Saudi Arabia or Iran?	`MATCH (a:Dataset)-[:HAS_LOCATION]->(l:Location) WHERE l.name CONTAINS 'Saudi Arabia' OR l.name CONTAINS 'Iran' RETURN a.name + " LINK: " + a.url AS response ORDER BY a.date DESC LIMIT 3`

3.3.3 Discussions

DataChat is a prototype conversational system for exploring dataset knowledge graphs, and is our initial step towards providing users with needed context and bridging gaps in stakeholder expertise through intelligent and user-friendly dataset search. DataChat leverages SKGs and LLMs to create a novel, interactive data search system. DataChat improves the four key aspects of research data search in CEVI. DataChat also provides a comprehensive research and development workflow from knowledge application design to end-user tool development, which is broadly applicable to data and digital curation applications.

Essential capabilities of SKGs to support dataset search, data reuse, and management: CEVI

DataChat enhances *connectivity* in ICPSR-SKG through linking entities previous disconnected, translating natural language input to Cypher queries, and integrating textual and visual information. These features benefit stakeholders (e.g., archivists, librarians, and bibliographers) by facilitating metadata management and dataset discovery [38], [75]. The improved *effectiveness* of DataChat, which replaces cross-tab search and multiple dropdowns with a single natural language input, makes the dataset search process useful for researchers, educators, and students, regardless of their technical expertise and time constraints. DataChat increases data *visibility* through graph visualization, which also highlights different attributes of nodes and the schema of ICPSR-SKG, enabling stakeholders to evaluate research impacts, identify gaps in knowledge, uncover potential collaborators, and gain insights into emerging research trends [51], [74]. Lastly, DataChat visualization's *interactivity* promotes user engagement by allowing users to emphasize specific nodes according to their needs and goals, creating a personalized experience as stakeholders explore research datasets.

LLMs bridge the human-database language gap, while performance varies by stakeholder

DataChat leverages GPT-3.5-turbo, one of the Generative Pre-trained Transformer (GPT) family's LLMs developed by OpenAI [256], known for their versatility in dealing with unseen scenarios or tasks which are essential abilities of artificial general intelligence. In general, LLMs support usability in SKG applications because they bridge the gap between natural language and graph database queries, enabling researchers to operate in network terms without prior knowledge about a specific type of database language. The GPT-3.5-turbo model works well for example inputs from education and data management unit stakeholders' perspectives. However, our evaluation indicated that the Cypher queries generated for stakeholders in the funding agency are not properly querying data from the ICPSR-SKG, possibly because of the complexity and ambiguity of those stakeholders' interests.

The Entity Ambiguation and Schema Alignment Challenges in the SKG-based Data Search

Integrating SKGs and LLMs has introduced advanced capabilities in academic data search. However, as the results in the DataChat prototype indicate, effectively implementing these technologies involves overcoming specific challenges related to entity ambiguation (for author name) and schema alignment (for keyword and hierarchical thesaurus). These challenges are critical to the system's accuracy and user efficacy.

Entity ambiguation, or the ability to correctly identify and track entities (authors) across publications, remains a significant problem. This challenge arises from the variability in how authors' names are presented across publications—variations in initials, last names, or the use of different names can lead to significant data integrity issues. Similarly, different authors with similar names can be mistakenly identified as the same individual. This misidentification can skew authorship data, affecting citation accuracy and research metrics. A robust disambiguation mechanism is essential for maintaining the reliability of authorship information within the SKG.

The alignment of schemas in the SKG, particularly regarding keyword and hierarchical thesaurus alignment, is crucial for enhancing the graph's functionality. First, accurate keyword alignment ensures that datasets and publications that are contextually related are connected within the SKG. This connection is vital for researchers seeking resources that align with their specific areas of inquiry, thus enhancing the graph's utility for discovering relevant research materials. Second, proper alignment within the hierarchical thesaurus is necessary to effectively map the complex and often multi-dimensional relationships between different research topics. This alignment aids in structuring the SKG for better navigability and utility, enabling users to explore interconnected research themes and datasets more efficiently.

Chapter 4

Constructing a Scholarly Knowledge Graph for Linked Research Data

The rapidly expanding volume of research data and its associated publications presents a significant challenge to academic researchers seeking efficient and effective ways to discover and reuse existing datasets [5], [6]. Traditional data management systems, which often rely on linear, metadata-driven search interfaces [9], [10], fall short in providing the nuanced understanding needed to navigate the complex relationships between datasets, metadata, and scholarly publications. In response to these challenges, SKGs have emerged as a transformative solution [49], offering a more dynamic and interconnected approach to data organization and searchability. Using SKGs, researchers can uncover latent connections and patterns that are not readily apparent using conventional search methods, thus enhancing the discoverability and reusability of research data [11], [49], [57]–[59].

As discussed in the prototyping reflections in Chapter 3, for example, in the DataChat prototype, the integration of the ICPSR Bibliography as a backend database illustrates the potential of a unified academic graph that includes both datasets and publications. However, reflecting on the current limited performance, it is imperative to further enrich and enhance metadata with well-curated external scholarly databases, for instance, OpenAlex [63], by incorporating diverse publications, extensive author networks, and, most importantly, high-quality entity disambiguation and schema alignment.

In this chapter[1], we discuss the construction of the ICPSR Health and Medical Scholarly Knowledge Graph (IHSKG). There are about 20% datasets in ICPSR that are about the health and medical domain, defined by its coverage of keywords 'health' and'medical' in the subject terms of the datasets, however, more than 40% of the publications in the ICPSR Bibliography cite these datasets. This high data usage rate in the health and medical domains underscores its utility and relevance as a sample for ICPSR, and more broadly, as a use case for institutional repositories.

As such, IHSKG helps to illustrate complex interrelations within medical and health research data. It also addresses potential algorithmic and infrastructure challenges. The DataChat prototype selects a period of time (2017-2019) to study data search while reducing infrastructural and engineering cost. Similarly to the mindset of obtaining a segment of the full data, IHSKG takes a domain-specific segment in the health and medical domain, which not only well presents the structure of the full ICPSR Bibliography but also being pertinent to

[1]The data source and related descriptions are used or derived from a co-authored dataset paper [257].

interdisciplinary data reuse. Although the processing time for graph database queries can increase with the size and complexity of data relationships, IHSKG remains a representative and essential subset of ICPSR. Figure 4.1 shows the IHSKG schema, illustrating the interconnectivity between datasets, publications, topics, and the various stakeholders such as authors, owners, and funders. Each type of entity in the graph is designed to include and express detailed metadata that enhance the discoverability and utility of scholarly data.

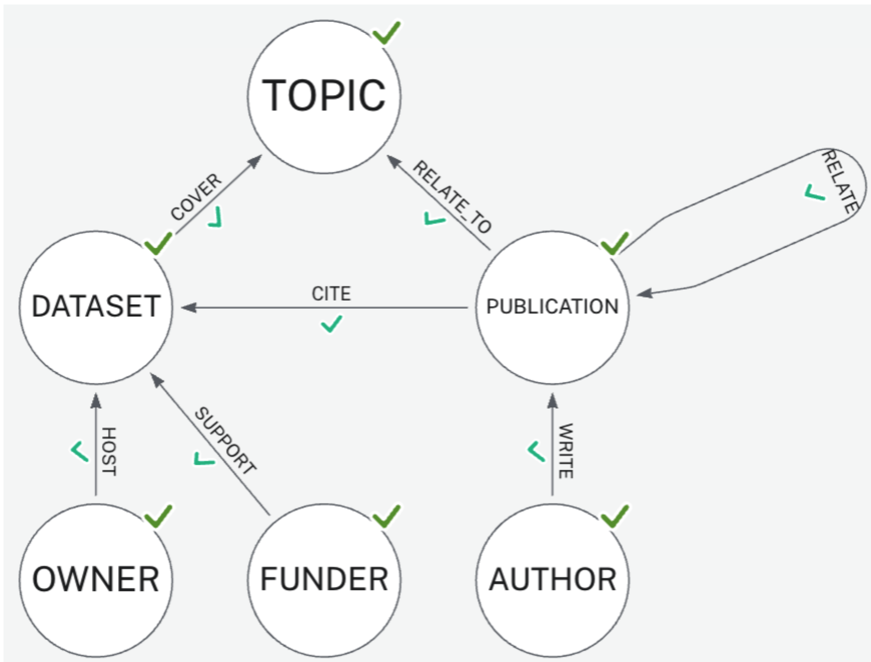

Figure 4.1: IHSKG Schema

Before we go into details of IHSKG's construction, we start with the introduction to the data sources in ICPSR, a primary resource for social science research datasets. We then focus on the infrastructure, schema and summary statistics of IHSKG, emphasizing its relevance for enabling robust, user-centric, and AI-augmented data search systems in the health and medical research domains.

4.1 Data Source

ICPSR provides access to more than 10,000 datasets and a bibliography of 100,000 data-related publications that have used those data. ICPSR makes several linked resources – including datasets, variables, and publications – available for search and discovery.

Curating the ICPSR Bibliography is labor intensive, and the current coverage of the ICPSR Bibliography is uneven. Bibliography staff search broadly for academic literature that references ICPSR studies and add literature to the Bibliography only if it analyzes ICPSR data or includes an extensive discussion of data-related methodology. Publications in the Bibliography are a mixture of materials published by the original data creator and publications that analyze data in ICPSR collections. The majority of materials are journal articles, reports, conference proceedings, theses, books, and book chapters.

Table 4.1 shows the data we used to construct the SKG, including their names, sources, and curation statuses.

Name	Source	Curation Note
Datasets (ICPSR Studies)	ICPSR Datasets	Done by ICPSR
Publications (ICPSR Papers)	ICPSR Bibliography	Done by ICPSR
Relations (Gold)	ICPSR Datasets Bibliography	Filtered by Author

Table 4.1: Data Source and Curation Notes

Studies (”ICPSR_STUDIES”): 10,684 social science research datasets available through ICPSR up to 2022-08-23 with variables for ICPSR study number, digital object identifier, study name, series number, series title, authoring entities, full text description, release date, funding agency, geographic coverage, subject terms, topical archive, curation level, single principal investigator (PI), institutional PI, total number of PIs, total variables in data files, question text availability, study variable indexing, level of restriction, total unique users downloading study data files and codebooks, total unique users downloading data only, and total unique papers citing data through August 2022. Studies map to the papers and curation logs table through ICPSR study numbers as “STUDY”. However, not every study in this table will have records in the papers and curation logs tables.

Papers (”ICPSR_PAPERS”): 107,862 unique publications collected from 2000-08-11 to 2022-08-23 in the ICPSR Bibliography and enriched with metadata from the Dimensions database with variables for paper number, identifier, title, authors, publication venue, item type, publication date, input date, ICPSR series numbers used in the paper, ICPSR study numbers used in the paper, the Dimension identifier, and the Dimensions link to the publication’s full text. Papers map to the studies table through ICPSR study numbers in the “STUDY_NUMS” field. Each record represents a single publication, and because a researcher can use multiple datasets when creating a publication, each record may list multiple studies or series.

Dataset-publication relations (Gold): The connection between the “ICPSR_STUDIES” and “ICPSR_PAPERS” is established through the “STUDY” and “STUDY_NUMS” fields, creating a many-to-many mapping. This means that a single study from the “ICPSR_STUDIES” table can be referenced in multiple publications within the “ICPSR_PAPERS” table. We regard the combined relations as a gold standard dataset of data-publication relations, signifying precise and accurate association between datasets and publications ensured by ICPSR’s professional curators. Figure 4.2 offers a holistic perspective on the intricate connections between datasets and publications and their associated metadata. The dataset is directly associated with the publication, indicating that a dataset is cited by a publication.

Table 4.2 further provides an example of available metadata for a highly cited ICPSR study.

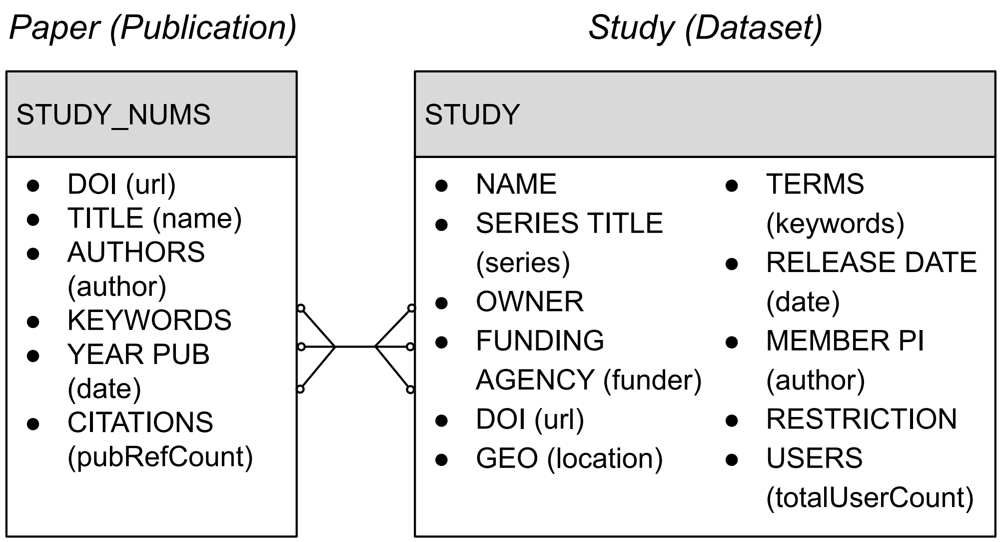

Figure 4.2: Selected ICPSR Metadata (before transformation)

Table 4.2: Example of available metadata for an ICPSR Study

Study name	Series title	Release	Citations	Subject Terms
Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 1996	Monitoring the Future (MTF) Public-Use Cross-Sectional Datasets	1998-10-05	251	attitudes, demographic characteristics, drug use, family life, high school students, life plans, lifestyles, social behavior, social change, values, youths

4.2 ICPSR Health and Medical Scholarly Knowledge Graph (IHSKG)

To advance the capabilities of the GAUDS system, we introduce the IHSKG. This innovative graph structures the metadata from the ICPSR's existing search system, focusing explicitly on the health and medical fields. Constructed using the Neo4j graph database platform AuraDB, the IHSKG capitalizes on Neo4j's robust support for natural language understanding and its ability to leverage intricate data connections. By encoding detailed information about the relationships among various research entities such as publications and datasets, the IHSKG facilitates dynamic interactions and enhances the contextuality of search results. This approach allows us to transcend the traditional limitations associated with static indexing systems, providing a more fluid and intuitive data exploration experience.

In this section, we will outline the selection criteria for including data and publication metadata within the IHSKG, focusing specifically on health and medical domains. These criteria are designed to ensure comprehensive coverage and relevance, allowing for the creation of interactive network visualizations. Through these visualizations, along with statistical and content analysis, we can achieve a deep understanding of large-scale relationships across various research entities. This holistic approach not only enriches the data search experience but also enhances the overall utility of the ICPSR's resources in supporting health and medical research.

Neo4j and AuraDB are the infrastructural foundations of the IHSKG, chosen for their robust handling of complex graph-based data structures. Neo4j's graph database management capabilities are well-suited for mapping the intricate relationships within ICPSR's scholarly data, because it is outstanding performance in recent empirical test [258] and is a popular choice in institutional repositories [259], [260]. Neo4j's cloud hosting platform, AuraDB, enhances this with cloud-based scalability and efficient large-scale data management, ensuring a stable and resilient environment for ICPSR's data needs, minimizing risks associated with data loss which could severely impact research continuity.

IHSKG is implemented on Neo4j Version 5 and currently encompasses 65,611 nodes and 183,011 relationships, occupying about 0.4 GB of storage. It operates under the free tier of Neo4j AuraDB, which, while cost-effective, imposes limitations on processing power and storage capacity. During periods of high demand, this may affect performance, potentially limiting the system's ability to conduct real-time data analysis. However, for the purposes of this prototype and the associated user study, this setup provides a suitable balance between functionality and cost.

Vector Indexing through Langchain is strategically utilized to refine data search capabilities within the IHSKG. The system establishes three distinct vector indices for dataset information: names, terms, and a combination of both. This allows for more precise searches that can identify datasets by name, by specific terms, or through a comprehensive search combining both elements. The decision to use OpenAI's 'text-embedding-3-small' model, which is a leading model in text-based vector search and optimized for latency and storage[67], for translating natural language queries into word vectors enables a nuanced understanding of user queries. This approach is critical in a research context, particularly in health and medical fields, where the accuracy and speed of data retrieval are paramount. By effectively mapping natural language into searchable vectors, the system not only speeds up the search process but also enhances the relevance of the results, providing researchers with quick access to vital information that can influence critical research decisions and healthcare outcomes.

4.2.1 Schema

The IHSKG schema effectively organizes various entity types and their relationships through nodes and edges, enhancing the system's ability to manage complex scholarly data. This schema is crucial for enabling comprehensive mapping and querying capabilities within the scholarly knowledge domain, particularly in the health and medical research fields. Below, we provide an in-depth overview of these components, complemented by Table 4.3 and Table 4.5, which list the attributes for nodes and the types of relationships for edges. We also discuss the usefulness of these structural elements and the reasons why we include them, emphasizing their importance for facilitating effective data discovery and reuse.

In this subsection, we delve deeper into the inclusion criteria for these nodes and the process of database transformation based on OpenAlex, illustrating how these elements contribute to the functionality of the

IHSKG. We first delve into the **Dataset** node and demonstrates the improved disocoverability through vector embeddings. We then discuss the inclusion of the **Publication** and **Topic** nodes, which were

Table 4.3: Attributes of Nodes in the IHSKG

Node Type	**Attributes**
Dataset	name, id_icpsr, terms, restriction, year, date, embedding, embedding_terms, embedding_name_terms
Publication	name, year, id_openalex, reference_count, related_works, datasets
Topic	subfield_id, subfield_name
Funder	name
Owner	name
Author	name, author_id

Dataset Node The attributes of dataset nodes, including *name*, *year*, and specific terms and embeddings, allow for nuanced filtering and identification processes. These attributes enhance the system's ability to perform targeted searches by reflecting both the basic metadata and the richer contextual details of the datasets. Through the similarity of dense word vectors via embedding, the use of specific and relevant terms may lead to similar search results, boosting the flexibility in user input and thus increase search efficiency and accuracy. For example, when using the word ”adolescence” to search in ICPSR's Find Data page, the top returned datasets have to contain the word stemmed from the search word, namely ”adolescents”, including datasets 22409, 22411, and 22410 (Figure 4.3); when using the GAUDS system and the same search word, the search results contain datasets with words such as ”young adults” (dataset 36012), ”childhood” (dataset 34870), and ”teenage” (dataset 6375) (Figure 4.4).

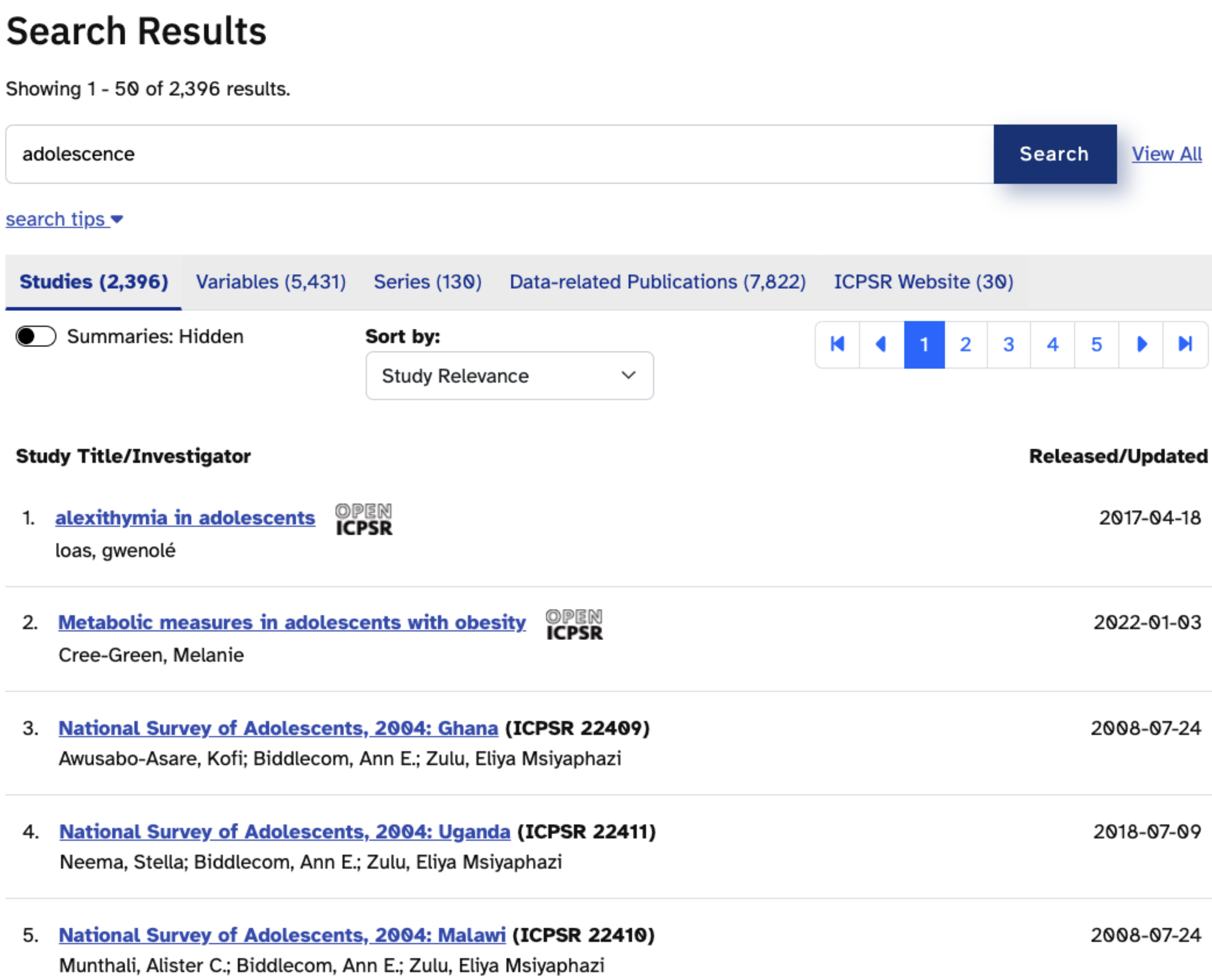

Figure 4.3: Dataset search example of ”adolescence” in the ICPSR find data page

DATA SEARCH RESULTS

Results based on query vector's similarity to datasets:


```
[
  0 : {
    "id" : 36012
    "name" : "Relationship Development and Health in Young Adults"
  }
  1 : {
    "id" : 36019
    "name" :
    "A Lifecourse Approach to Emerging Health Disparities in a US Birth Cohort"
  }
  2 : {
    "id" : 34870
    "name" :
    "Development and Malleability from Childhood to Adulthood in Baltimore,
    Maryland, 2001-2005"
  }
  3 : {
    "id" : 35952
    "name" :
    "Daily Experience in Adolescence and Biomarkers of Early Risk for Adult
    Health"
  }
  4 : {
    "id" : 6375
    "name" : "Teenage Attitudes and Practices Survey II, 1993: [United States]"
  }
  5 : {
    "id" : 35249
    "name" :
    "National Longitudinal Study of Adolescent to Adult Health (Add Health),
    1994-2008 [Restricted Use]"
  }
```


Figure 4.4: Dataset search example of ”adolescence” in the GAUDS vector search

Publication Node By linking publications with datasets and detailing their interactions through reference counts and related works, this node enables researchers to trace the scholarly impact and assess the relevance of publications to their ongoing projects. Such connectivity also facilitates the assessment of data usage across different studies, in particular, through co-citation analysis of datasets. For example, co-citation networks can

Topic Node Incorporating nodes for specific subfields aids in categorizing datasets and publications into well-defined topics, simplifying targeted searches and thematic analysis throughout the database. This structural choice supports detailed academic inquiries and enhances the accessibility of thematically relevant data. Supported by OpenAlex's updated topic classification system [63], even if a data user is not familiar with the formal name of a topic, they can still search with the keywords they are familiar with while the GAUDS system can provide useful classification of dataset topics in need and point the user to the frequently used datasets under the topics. For example, if we use the word ”adolescence” as the query, there are three suggested topics, including ”Reproductive Medicine,” ”Pediatrics, Perinatology and Child Health,” and ”Psychiatry and Mental health” (Figure 4.5). In addition, the overview of top topics in IHSKG reveal the dataset landscape and archiving focus in the health and medical domains in ICPSR. Table 4.4 provide the top topics of the datasets, which are helpful to curators and funders who need to assess and improve data work. The distribution aligns well with the social science nature of ICPSR, where the ”Public Health, Environmental and Occupational Health” topic is the most frequent with 1915 counts, which is more than two times the second frequent topic and nearly five times the third frequent topic.

DATA SEARCH RESULTS

Query:

adolescence

Topics:


```
[
  0 : "Reproductive Medicine"
  1 : "Pediatrics, Perinatology and Child Health"
  2 : "Psychiatry and Mental health"
]
```


Top datasets by topic:


```
[
  0 : {
    "d.name" : "Demographic and health Surveys STATcompiler"
    "d.id_icpsr" : 144
  }
  1 : {
    "d.name" : "GenderStats"
    "d.id_icpsr" : 146
  }
  2 : {
    "d.name" : "Cebu Longitudinal Health and Nutrition Survey"
    "d.id_icpsr" : 178
  }
  3 : {
    "d.name" : "Fragile Families and Child Wellbeing Study"
    "d.id_icpsr" : 180
  }
```


Figure 4.5: Dataset search example of ”adolescence” in the GAUDS topic search

Table 4.4: Dataset Distribution across Topics (Subfields) in IHSKG

Topic ID	Topic Name	Dataset Count
3600	Public Health, Environmental and Occupational Health	1915
2739	Radiological and Ultrasound Technology	813
2713	Physical Therapy, Sports Therapy and Rehabilitation	405
3616	Orthopedics and Sports Medicine	279
2735	Geriatrics and Gerontology	278
3605	Pediatrics, Perinatology and Child Health	169
2916	Transplantation	135
2711	Dermatology	124
2910	Urology	80
2725	Epidemiology	79

Funder, Owner, and Author Nodes These nodes outline the broader ecosystem surrounding datasets and publications by identifying funders, owners, and authors (of publications). Such transparency not only fosters accountability but also encourages collaboration within the research community by making visible the networks of funding and authorship.

Based on the nodes, we also include a set of relevant edges as the relations among nodes. Table 4.5 shows the relationship type, the source (from which node), the target (to which node), and the detailed descriptions of the links. These links further support the discovery and connectivity of the nodes, enabling scientific discoveries through relations such as co-citations, which can identify more possible datasets to use [261], and related publications, which enlarge the size of dataset use case references and provide more examples for informed reuse [262], [263].

Table 4.5: Relationships between Nodes in the IHSKG

Relationship Type	Source	Target	Description
COVER	Dataset	Topic	Indicates a dataset's coverage of specific topics
CITE	Publication	Dataset	Shows where a publication has cited a dataset
RELATE_TO	Publication	Topic	Connects publications to the topics they relate to
RELATE	Publication	Publication	Identifies relations between publications
SUPPORT	Funder	Dataset	Links funders to the datasets they support
HOST	Owner	Dataset	Indicates the ownership and stewardship of datasets
WRITE	Author	Publication	Connects authors to the publications they have written

4.2.2 Summary Statistics and Case Study

The IHSKG provides a new paradigm for analyzing the reuse and impact of datasets within health and medical research. This knowledge graph not only facilitates the understanding of why certain datasets are highly cited but also illuminates their contribution to advancing the field. Additionally, data users can benefit from accessible statistics of datasets, publications, and other metadata, which enhance the dataset discovery and reuse experience.

Summary Statistics of IHSKG

The IHSKG contains a rich array of nodes and edges, each representing crucial elements of scholarly communication and dataset lifecycle. Table 4.6 a breakdown of the number of nodes and edges within the knowledge graph.

Table 4.6: Summary of Nodes and Edges in the IHSKG

Entity Type / Relationship	**Count**	**Description**
Authors	38,269	Total number of authors in the database
Publications	24,890	Total number of publications in the database
Topics	63	Total number of unique topics covered
Owners	19	Total number of dataset owners
Funders	181	Total number of funders
Datasets	2,189	Total number of datasets
Host	7	Links owners to the datasets they host
Relate To	26,819	Connects publications to related topics
Support	2,499	Links funders to the datasets they support
Cite	47,777	Shows publications that cite datasets
Relate	5,272	Identifies relations between publications
Write	93,624	Connects authors to the publications they write
Cover	4,831	Indicates datasets' coverage of specific topics

The IHSKG also obtained a density of $8.5 \times e^5$, higher than the original graph's density of $2.3 \times e^5$ [261]. These statistics highlighting the cohesiveness and extensive interconnectedness of datasets, publications, and other scholarly entities within the IHSKG. These statistics demonstrating the graph's capability to provide comprehensive insights into the academic landscape of health and medical research.

Analysis of Dataset Reuse in IHSKG with a Case Study

Based on IHSKG, we can analyze dataset reuse from several different dimensions, such as citation patterns, among many other possible research directions. Analyzing the citation patterns of datasets helps identify which datasets are highly cited and potentially why. This can reveal trends in dataset popularity and importance in the field. We can discover and analyze top cited datasets in the IHSKG, providing a clear perspective on which datasets are most influential in health and medical research. For instance, Table 4.7 includes the dataset name, its ICPSR ID, and the number of citations it has received for the top 10 most cited datasets in IHSKG. Among all datasets, for example, the ”National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2018” dataset is the most frequently cited datasets with a citation count of 2544, which is more than two times of the second most cited dataset. This observation shows a significant difference of reuse frequency among different datasets in IHSKG. This initial insight can provide further research and data management directions, where health and medical researcher can reuse datasets with thorough previous use cases and data management units can measure the impact of datasets.

Table 4.7: Top Cited Datasets in the IHSKG

Dataset Name	**ICPSR ID**	**Citations**
National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2018 [Public Use]	21600	2544
National Health and Nutrition Examination Survey III, 1988-1994	2231	1237
Midlife in the United States (MIDUS 2), 2004-2006	4652	989
National Health and Nutrition Examination Survey (NHANES), 2001-2002	25502	742
Collaborative Psychiatric Epidemiology Surveys (CPES), 2001-2003 [United States]	20240	718
National Health and Nutrition Examination Survey (NHANES), 1999-2000	25501	693
National Health and Nutrition Examination Survey (NHANES), 2003-2004	25503	624
Midlife in the United States (MIDUS 1), 1995-1996	2760	580
National Longitudinal Surveys of Labor Market Experience, 1966-1992	7610	497
NICHD Study of Early Child Care and Youth Development: Phase I, 1991-1994 [United States]	21940	466

We also list several other potential use cases, including topic coverage, funding and support, and collaborative networks:

- Topic Coverage: Examining the topics covered by highly reused datasets can provide insights into research areas that are currently active or emerging. This helps in understanding how datasets contribute to advancing specific areas of medical research.

- Funding and Support: By studying the support edge, we can identify how funding influences dataset creation and reuse. This may highlight the role of specific funders in promoting high-impact research.

- Collaborative Networks: Analyzing the 'Write' and 'Relate' edges can shed light on collaborative networks within the field. This includes examining how authors collaborate on publications and how publications interlink, which can influence the reuse of datasets.

4.2.3 Discussions

Inclusion Criteria for Nodes

The design of the IHSKG is guided by specific inclusion criteria for nodes that are critical for maintaining a focused and efficient database. These criteria are devised to enhance the utility and relevance of the data within the SKG.

Useful for Dataset Reuse Search Nodes included in the IHSKG must directly facilitate or enhance the searchability and reuse of datasets. This criterion leads to the exclusion of internal curation notes, which are typically used for data management and administration purposes but do not provide value to external researchers. By focusing on externally relevant metadata, the graph remains streamlined and more accessible for users seeking specific data for their research. The reliance on controlled metadata ensures that the included nodes are consistently described, improving the accuracy and efficiency of dataset searches. Standardized metadata schemas provided by ICPSR and expanded with OpenAlex [264] are employed to maintain high-quality data description, which is crucial for effective data discovery and reuse.

Focus on Stakeholders Nodes must be relevant to the stakeholders of the research community, which include researchers, funders, publishers, and academic institutions. This focus ensures that the information within the knowledge graph serves the practical needs of these users. Consequently, information that does not directly affect stakeholder decisions or research, such as series identifiers and geographic locations (unless they are significant to the research), are excluded. This prioritization helps in tailoring the graph to support meaningful research connections and data discovery. The use of well-defined metadata schemas supports this goal by providing a clear and structured way to represent information, making it easier for stakeholders to find and use the data they need. This approach underscores the importance of adopting metadata schemas that are widely recognized and effective for data discovery, ensuring the IHSKG remains a valuable resource for its users.

OpenAlex-based Entity Disambiguation, Alignment, and Extension

OpenAlex [264] is a scholarly database that offers a rich set of tools and data for disambiguating authors, aligning keywords and hierarchies, and extending bibliographic and metadata entries with related works. OpenAlex's capabilities are instrumental in enhancing the accuracy and breadth of the IHSKG.

Author Disambiguation. One of the significant challenges in managing scholarly databases is the issue of author disambiguation. OpenAlex provides robust methods for identifying and distinguishing between authors who may share similar names but are distinct individuals. This capability is crucial for ensuring that publications and datasets are accurately attributed to the right researchers, enhancing the reliability of the data within the IHSKG.

Keyword and Hierarchy Alignment. OpenAlex facilitates the alignment of keywords and the establishment of hierarchical relationships within the topics and fields of study. This alignment ensures that the vocabulary used across the IHSKG is consistent and accurately reflects the current understanding and organization of knowledge domains. Such structured taxonomy is vital for effective data querying and retrieval. Here, we use word-embedding-based alignment to map the terms of each dataset to the top three relevant subfields, renamed as topics in IHSKG, from the OpenAlex database.

Extension with Related Works. By incorporating OpenAlex's data on related works, the IHSKG can extend its nodes to include linked scholarly articles, datasets, and other pertinent works. This extension not only enriches the nodes with additional contextual information but also creates a more interconnected and comprehensive network of scholarly resources. This feature is particularly valuable for researchers looking to explore the breadth of literature and data related to their specific interests.

Enhanced by OpenAlex, the IHSKG achieves a robust level of disambiguation, alignment, and extension, making it a comprehensive and reliable resource for the research community. This integration ensures that the knowledge graph remains current and aligned with the latest scholarly communications and metadata standards, thereby enhancing its functionality and value to users.

Topics versus Keywords in Representing Research Themes

In scholarly databases like the ICPSR and OpenAlex, the distinction between topics and keywords becomes crucial for enhancing data discoverability and usability. This distinction impacts how users interact with and benefit from such systems. Below, we delve into how both ICPSR and OpenAlex approach the organization and classification of scholarly content, highlighting the potential implications for research and data discovery.

ICPSR emphasizes the use of both keywords and structured topic classifications to improve the discoverability of datasets within its vast archives. According to a study on search strategies at ICPSR, users commonly employ keyword-based queries that incorporate terms related to the subject, geographic location, or time period relevant to their research interests. However, these keyword searches often do not utilize the rich metadata and structured topics available, which could enhance the precision and context of search results.

ICPSR recognizes the need for a system that supports both exploratory and directed searches [56]. Exploratory searches often involve broad keyword queries that help users learn about and navigate through different research domains, while directed searches are more precise, utilizing specific dataset names, series, or authors. The challenge lies in aligning user-generated keywords with the structured topics and metadata within the ICPSR database, ensuring that searches are both intuitive and powerful enough to leverage the full scope of the repository's resources.

OpenAlex approaches the topic classification with a robust system designed to succeed where the previous Microsoft Academic Graph (MAG) [180] system may have faltered. Initially, OpenAlex adopted MAG's field of study data to ensure continuity and familiarity for users transitioning between systems. However, the

MAG system was critiqued for its convoluted concept hierarchy, issues with term polysemy and ambiguity, and the static nature of its topics, which failed to evolve alongside shifting academic landscapes.

To address these challenges, OpenAlex developed a more dynamic and granular topic classification system using a deep learning model informed by the CWTS classifications, which assigns ”micro-level” fields to each paper based on citation data [63]. This model not only adjusts to new trends but also enhances granularity by applying topics at the paper level rather than the journal level, using established structures like Scopus's ASJC codes. This system allows for a more precise and relevant categorization of papers, facilitating better match between user queries and scholarly content.

Therefore, in the construction of IHSKG, we recognize the impact of choosing topics versus keywords for scholarly research and data discovery. The evolution of topic classifications in SKG can have significant implications for research efficiency and data discoverability in several different aspects. First, topics can provide improved search accuracy. By employing more sophisticated and contextually aware topic models, systems like ICPSR and OpenAlex can improve the accuracy of search results, reducing the time researchers spend sifting through irrelevant data. Second, topics can enhance interdisciplinary collaboration. More intuitive and granular topic classifications help bridge disciplinary divides, enabling researchers from different fields to find and collaborate on cross-disciplinary projects more effectively. Thrid, topics can lead to adaptive learning systems. As systems evolve to incorporate user feedback and real-time data trends, they become more adept at predicting and accommodating researchers' needs, potentially transforming how knowledge is discovered and consumed. Such customizable and flexible mechanism can enable better resource allocation. For instance, enhanced data discoverability can influence funding decisions and policy making, directing resources to topics (areas) with high engagement and potential for impact.

In conclusion, the strategic application of topics versus keywords in scholarly knowledge bases offers a promising path toward more intuitive, efficient, and comprehensive research tools. Both ICPSR and OpenAlex are pioneering approaches that could serve as models for future developments in scholarly data management and retrieval systems.

Chapter 5

Assessing Large Language Models' Abilities for Complex Reasoning

As scholarly research becomes increasingly data-driven and complex, the need for sophisticated tools to process and analyze vast amounts of information has become critical. LLMs have shown great promise in addressing these needs by offering advanced capabilities in natural language processing and complex reasoning. This chapter explores the critical task of selecting the right LLM for specific scholarly tasks, focusing particularly on their ability to perform complex reasoning and learning in the complex reasoning scenario, addressing and preparing for the reasoning needs in understanding data structure and customizing models for SKGs.

SKGs, integrated with technologies such as LLMs and vector search, provide a robust framework for developing advanced data discovery platforms [42], [43]. These platforms harness the semantic search capabilities of LLMs to interpret and respond to user queries with unprecedented relevance and contextual awareness, making them indispensable in multidisciplinary research environments [67]–[69]. The ability of LLMs to synthesize and cross-reference information across diverse fields can significantly enhance the research process, offering more intuitive and effective search and data analysis tools [50], [51], [63].

Selecting LLMs for handling complex tasks[1] can be challenging. Choosing the appropriate LLM for specific tasks involves understanding their capabilities in reasoning and learning, especially when these tasks involve complex, logical problem solving such as constructing and querying SKGs. The selection process is crucial as the complexity of the tasks, such as generating Cypher queries for graph databases, requires a model not only with high reasoning capabilities but also one that can adapt and learn from new contexts.

Traditional benchmarks for evaluating LLMs, such as MMLU [266] and GAOKAO [267], focus largely on human-generated questions and standard answers, which do not fully test the limits of LLMs in complex reasoning scenarios. Benchmarks like Big-Bench Hard [268], DROP [269], and HellaSwag [270], while valuable, primarily assess multi-step reasoning, reading comprehension, and commonsense reasoning, but do not prioritize complex logical reasoning. This chapter introduces **NPHardEval**, a benchmark designed to rigorously test the logical reasoning capabilities of LLMs by emphasizing dynamic, logic-based reasoning challenges that mirror the complexities encountered in real-world data analysis [271]. NPHardEval goes

[1]Most of the content in this chapter is derived from a paper co-authored by the author, accepted to the ACL 2024 main conference [265].

beyond conventional benchmarks by incorporating tasks that require a deep understanding of complex reasoning, addressing the need for a robust and quantitative assessment of LLMs. This benchmark evaluates LLMs based on their performance in dynamic logic-based reasoning scenarios, filling a crucial gap in existing methodologies.

The rest of the chapter will detail the construction of NPHardEval, including its definition of complexity classes, designation of task difficulty levels, the synthesis of data, the evaluation metrics, and the models to evaluate. Additionally, it will explore the LLMs' reasoning abilities through performance comparisons and their learning capabilities through in-context learning and fine-tuning strategies. This approach not only ensures that the selected LLM candidates are optimally suited for the demands of sophisticated data management systems such as the GAUDS system, but also pushes the boundaries of what these models can achieve in terms of data analysis and interpretation in academic research.

5.1 NPHardEval Benchmark Construction

In this section, we discuss the construction of the NPHardEval benchmark, which serves as a generalized representation of complex scenarios in the GAUDS system. The reasoning needs of questions in the NPHardEval Benchmark mimic the challenges we expect to face in the in the GAUDS system. The evaluation results of LLMs on this benchmark can shed light on the expected performances of candidate models of tasks associated with different features in GAUDS.

5.1.1 Complexity Classes

In our study, we employ the concept of complexity classes to categorize the reasoning tasks for LLMs. These classes are defined based on computational resources, such as time or memory, required to solve the problems they contain [272]. Primarily, most complexity classes comprise decision problems that can be solved using a Turing machine, with differentiation based on their time or space (memory) requirements. For example, class P includes decision problems that a deterministic Turing machine can solve in polynomial time. Tasks within this class often pose multidimensional cognitive challenges, enriching the evaluation framework of LLMs. This structured approach not only aids in assessing the reasoning capabilities of LLMs but also holds substantial relevance in various practical applications, particularly in the optimization and decision-making process.

In particular, we use three complexity classes to define the task complexity in the benchmark, including P (polynomial time), NP-complete (nondeterministic polynomial-time complete), and NP-hard, which are increasingly complex in both the intrinsic difficulty and the resources needed to solve them. Figure 5.1 shows their relation regarding computational complexity in a Euler diagram. The details of the nine problems, including the Graph Coloring Problem Optimization Version (GCP), Traveling Salesman Problem Optimization Version (TSP), Meeting Scheduling Problem (MSP), Knapsack Problem (KSP), Traveling Salesman Problem Decision Version (TSP-D), Graph Coloring Problem Decision Version (GCP-D), Shortest Path Problem (SPP), Edit Distance Problem (EDP), and Sorted Array Search (SAS), are provided in Appendix A.2. This approach aims to delineate the extent of complex reasoning achievable by LLMs, thus, for each complexity class, we only choose tasks from the nonoverlapping subset of the complexity class. In our selection criteria, we intentionally exclude tasks that require intensive mathematical computations, such as matrix multiplication and logarithmic calculations. Thus, we do not list the NP class (questions in NP but not P and not NP-complete), which is exemplified by the discrete logarithm and integer factorization problems, as the majority of such problems are characterized by their calculation-intensive nature (see details in the Appendix A.2).

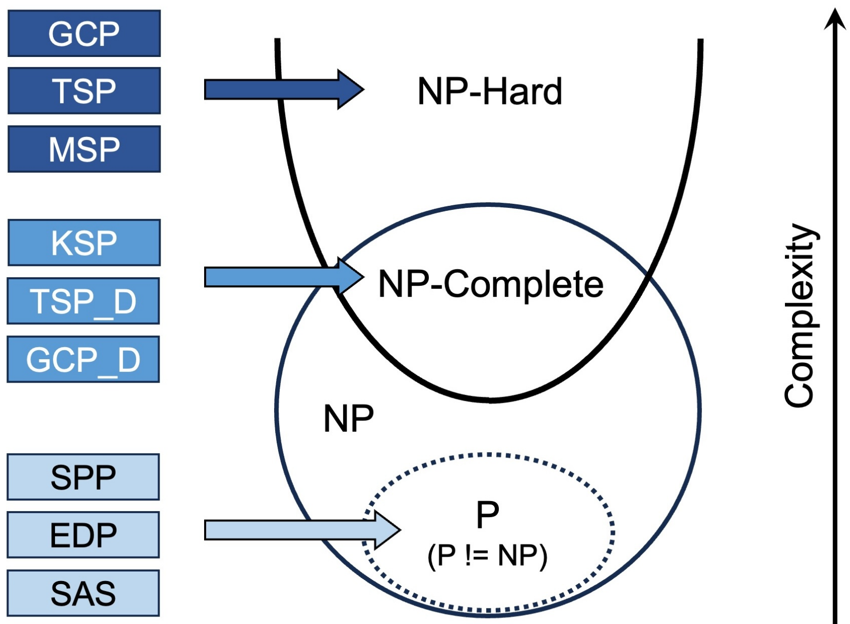

Figure 5.1: Complexity of Benchmark Questions.

5.1.2 Difficulty Level for Tasks

NPHardEval categorizes each challenge into a hierarchy of difficulty, spanning from the simplest to the most difficult with 10 levels. This gradation allows for a nuanced assessment of an LLM's problem-solving abilities across a spectrum of increasingly difficult tasks. For instance, the GCP-D problem has difficulty levels 1 to 10 with questions of 6, 8, 10, 12, 14, 16, 18, 20, 22, and 24 average edges and 6, 7, 8, 9, 10, 11, 12, 13, 14, and 15 nodes. Beginning with graphs of 6 nodes and 6 edges, each subsequent level incorporates an additional 2 edges and 1 node, culminating in graphs of 24 edges and 15 nodes.

The difficulty level is not strictly bounded to a linear scaling of difficulty; rather, it is designed to explore the nuances of performance degradation. By observing how LLMs cope with an escalating series of challenges, we aim to identify the inflection point where the performance notably diminishes. This approach provides a comprehensive understanding of where LLMs excel and where they fail, forming potential pathways to enhance their reasoning capabilities.

5.1.3 Data Synthesis

In the context of data synthesis for complex tasks, the approach can be categorized into two distinct methodologies, each corresponding to a different type of data structure: graph data (e.g., GCP) and linear data (e.g., MSP). The synthesis process in both cases is governed by a progression of complexity across a spectrum of predefined levels. This structured approach enables the creation of diverse datasets, suitable for evaluating and benchmarking LLMs' reasoning ability. We provide examples of the synthesized data and how they are used in the prompts in Appendix A.1.

Graph Data Synthesis The complexity in graph data synthesis escalates through a series of levels, each defined by a set of parameters that dictate the graph's size and intricacy. These parameters typically include the number of vertices, the number of edges, and the range of edge weights. At lower levels, graphs are simpler, with fewer vertices and edges and a limited range of edge weights. As the level increases, the graphs become progressively more complex, featuring more vertices, a higher density of edges, and a wider variety of edge weights. The synthesis process is as follows:

- A generative function is employed to construct individual graph instances. This function adheres to the principles of graph theory, ensuring the creation of simple graphs without self-loops and duplicate edges, and respecting the parameters dictated by the current difficulty level.
- A batch synthesis function then iteratively employs the generative function to produce multiple graph instances across the spectrum of difficulty levels.
- Finally, the synthesized graph instances are preserved in a tabulated format (in a CSV file), facilitating subsequent utilization and analysis.

Linear Data Synthesis In linear data synthesis, complexity is modulated by manipulating the length of the data array and the range of its constituent elements. Initial levels are characterized by shorter arrays with elements drawn from a narrow range. As the difficulty level increases, the arrays lengthen, and the range of possible element values expands, thus introducing greater variability and complexity to the problem. The synthesis process is as follows:

- A linear data instance generation function is first utilized. This function produces sorted arrays of random numbers within a defined range and selects a target number, ensuring its presence within the array to guarantee solvability.
- Multiple instances are generated through an iterative process, adhering to the difficulty levels outlined.
- These instances are then systematically recorded in a structured format (in a JSON file) for easy access and analysis.

5.1.4 Evaluation Metrics

To evaluate the reasoning ability of LLMs, we utilize two metrics, the Weighted Accuracy and the Failure Rate, to comprehensively quantify the correctness of LLMs' reasoning output.

Weighted Accuracy (WA) is calculated for each problem through a comparison with the correct answer or through a step-by-step results check, for those problems without the only answer. To better represent the comparative accuracy, we assign weights to different difficulty levels so that each level has a weight corresponding to its relative importance or challenge. Higher difficulty levels are given more weight in a linear manner (e.g., level 1 has weight 1, level 2 has weight 2, etc.). The Weighted Accuracy is defined as:

$$WA = \frac{\sum_{i=1}^{10}(w_i \times A_i)}{\sum_{i=1} 10} w_i$$

where w_i represents the weight assigned to the difficulty level i, from 1 to 10, and A_i is the accuracy at that level.

Failure Rate (FR) is a measure used to assess the frequency of unsuccessful outcomes in different problems and difficulty levels. It is particularly useful for identifying cases where an LLM's result does not comply with the expected output format. The Failure Rate is calculated by considering the proportion of failed attempts relative to the total number of attempts for each difficulty level. An attempt is defined as failed if the model generates results that cannot be successfully parsed in all endpoint calls, and we set the maximum times of try as 10. For each problem, the Failure Rate is then aggregated across all difficulty levels, taking into account the total of 10 attempts at each level. The formal definition of Failure Rate is given by:

$$FR = \frac{\sum_{i=1}^{10} F_i}{100}$$

where F_i denotes the number of failed attempts at difficulty level i.

5.1.5 Models to Evaluate

We evaluate 12 LLMs including closed-source models (GPT 4 Turbo [273], Claude 2 [274], GPT 3.5 Turbo [275], Claude Instant [276], and PaLM 2 [277]) and open-source models (Yi-34b [278], Qwen-14b [279], Mistral-7b [280], Phi-2 [281], MPT-30b [282], Vicuna-13b [283], and Phi-1.5 [284]) across three complexity classes (P, NP-complete, NP-hard) each with 10 difficulty levels.

This comparison sheds light on the relative strength and weakness of these models and determines their proficiency in solving progressively challenging problems, thus gauging their ability to handle tasks with increasing complexity.

5.2 Reasoning Ability

LLMs [285]–[287] have made significant advances in natural language processing and related fields. Recent research underscores the unprecedented reasoning abilities of LLMs in various fields, from biomedical and human-computer interaction research to humanities and social studies [288]–[292]. It has been discussed that these models exhibit "emergent" behaviors, including the ability to "reason" when they are large enough [293], [294]. By providing the models with the chain of thoughts with a simple prompt "Let us think step by step", these models are able to answer questions with explicit reasoning steps [295]. This has sparked considerable interest in the community since reasoning ability is a hallmark of human intelligence. Various variations of chain-of-thought have been developed to prompt models' reasoning ability [296]–[298], such as tree of thought [299], graph of thought [300], and self-inspring technique [301].

Later, various self-critique methods have been proposed to enhance LLM's reasoning performance. The Recursively Criticizes and Improves (RCI) approach, for example, iteratively refines output, proving to be more effective in automating computer tasks and elevating reasoning capabilities [302]. Backward verification proposes an intuitive human-like mechanism for LLMs to self-check and improve their conclusions, reducing errors in reasoning tasks [303].

5.2.1 Experiment: Reasoning Performance Comparison

To evaluate the reasoning abilities of different LLMs through the **NPHardEval** benchmark, we employ a comparative experimental design. We use zero-shot prompts containing task descriptions and a specific question as the foundational measure of performance. The performance of each model was evaluated on the basis of two primary metrics: weighted accuracy and failure rate across the different complexity classes of problems. We also conduct few-shot experiments on SAS and EDP using GPT-4 generated examples in prompts and observe interesting performance change across different types of prompts.

To evaluate across task complexity, specifically comparing the complexity among pairs P, NP-Complete, and NP-Hard, we initially pinned the data based on complexity levels. Subsequently, we applied the Wilcoxon test to each pair of complexity sets. Wilcoxon is a non-parametric statistical hypothesis test that allows us to compare two populations with matched samples. To evaluate problem difficulty, with the aim of discerning differences between problems within the complexity category, we pinned the data on the specific problems and then used the Wilcoxon test to compare pairs of different problem sets.

5.2.2 Results: Reasoning Ability of LLMs

The reasoning ability experiment focuses on a comprehensive comparison among various foundation models and across complexity classes and difficulty levels. In Figure 5.2, we present the overall zero-shot accuracy for each problem, providing a visual representation of the performance of different models.

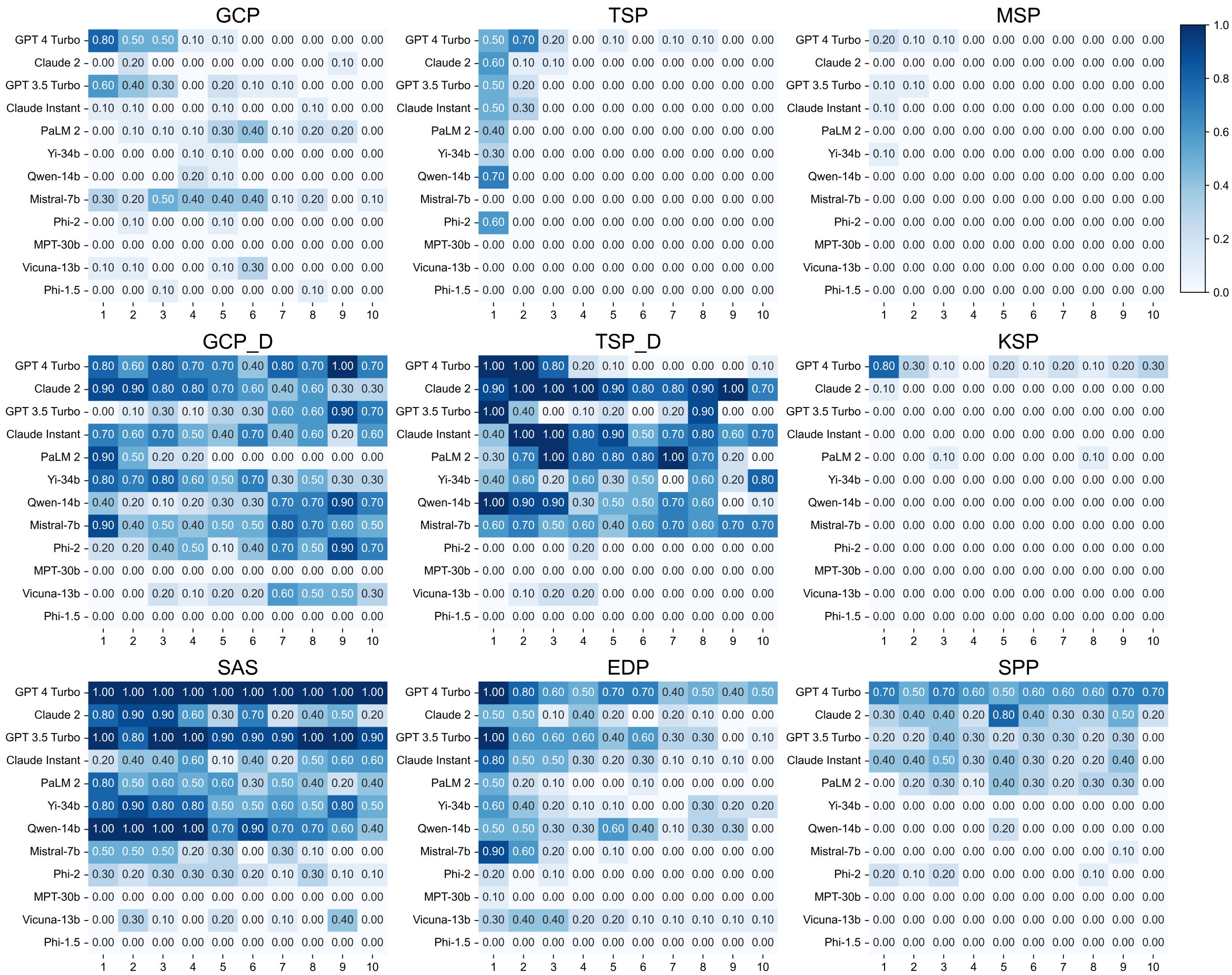

Figure 5.2: Zero-shot model performance on the nine tasks from P to NP-Complete bottom-up.

Our observations reveal that closed-source models generally demonstrate higher accuracy and a lower rate of failure compared to their open-source counterparts. Notably, GPT-4 Turbo often emerges as the frontrunner in performance across the majority of tasks, indicating its superior problem-solving capabilities, while Claude 2, on the other hand, often performs the best on medium-level (NP-complete) complexity in zero-shot settings. Within the realm of open-source models, Yi-34b, Qwen-14b, and Mistral-7b distinguish themselves by significantly outperforming other models in this category. We observe a disparity between the performance of these three models and other open-source options, highlighting a notable performance gap and suggesting that these models possess more advanced reasoning abilities.

In particular, we used the weighted accuracy and the failure rate metrics to further quantify different models' performance. The trends observed below in both weighted accuracy and failure rates point to a nuanced understanding of the capabilities and limitations of current LLMs. These observations are also supported by statistical tests within and across complexity classes, indicating the model performance differences among complexity classes.

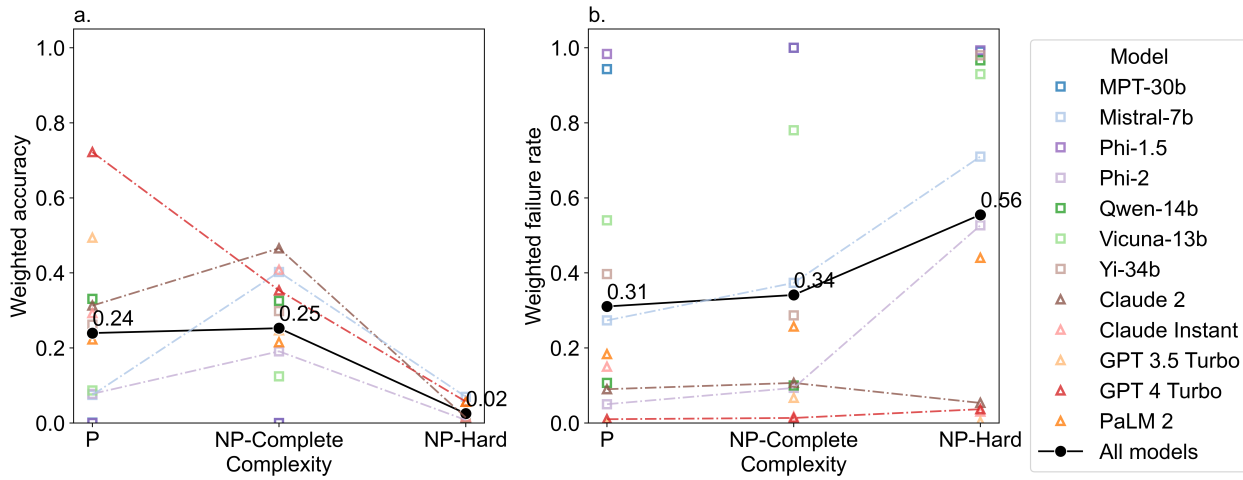

Figure 5.3: Model performance on different complexity problems: (a) weighted accuracy (b) (weighted) failure rate. Open models are denoted in squares and close models are denoted in triangles. Trends of metrics are demonstrated for models with outstanding performances in both weighted accuracy and failure rate, including both close-source (GPT 4 Turbo and Claude 2) and open-source (Mistral-7B and Phi-2) models.

Weighted Accuracy Figure 5.3(a) shows the weighted accuracy for different models across problem complexities. The general trend is all models experiencing a decrease in accuracy as problem complexity increased. Notably, there are two detailed findings for overall reasoning ability change. First, regarding the performance decay speed, among the 12 models we tested, the average performance demonstrated a higher accuracy at the P and NP-Complete complexity levels (with similar weighted accuracies of 0.24 and 0.25) but saw a sharper decline as the problems became more complex when proceeding to the NP-hard level (with a weighted accuracy of 0.02). There is a *performance decay* on average when models are tested against NP-Hard problems. Second, close-source models usually perform better than open-source models: there are more triangles in the upper locations than squares in Figure 5.3(a).

Failure Rate Figure 5.3(b) indicates that the failure rates mirrored the trends observed in weighted accuracy but in reverse. On average, the models showed an increase in failure rates corresponding to the complexity of the problems. Open-source models fail more often (with more squares on the top) than the close-source models (with more triangles on the bottom), indicating close-source's models advanced ability in following the prompt to understand the reasoning problems and generate answers with correct format.

5.2.3 Discussion: Performance of the Model at Task Complexity and Difficulty Levels

Across Complexity Levels

Figure 5.4 shows the accuracy of each model at different complexity levels. The test results reveal statistical significance ($p < 0.05$) in the p values between P and NP-Hard, as well as NP-Complete and NP-Hard. These findings indicate that our investigated LLMs performed significantly worse when faced with NP-Hard problems compared to P and NP-Complete problems.

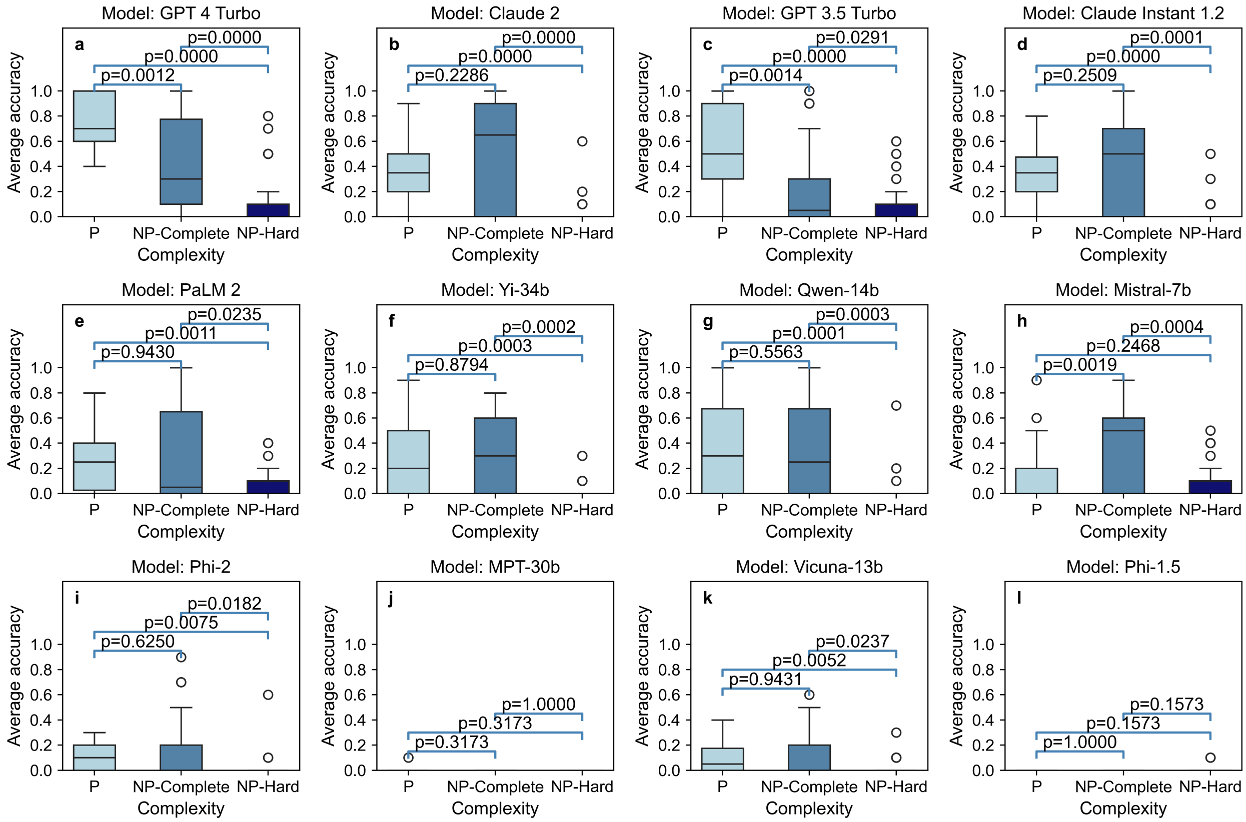

Figure 5.4: Models' performance on each complexity level. (a) GPT 4 Turbo. (b) Claude 2. (c) GPT 3.5 Turbo. (d) Claude Instant 1.2. (e) PaLM 2. (f) Yi-34b. (g) Qwen-14b. (h) Mistral-7b. (i) Phi-2. (j) MPT-30b. (k) Vicuna-13b. (l) Phi-1.5.

Across Models

Figure 5.5 presents the accuracy of each model in various problems associated with the complexities of P, NP-complete and NP-hard. Regarding P complexity, notable differences emerged among the models. GPT 3.5 Turbo, GPT 4 Turbo, Yi-34b, and Qwen-14b models exhibited significantly superior performance on the SAS problem compared to the other two problems. GPT 3.5 Turbo, Yi-34b, and Vicuna-13b models demonstrated markedly better performance on the EDP problem compared to the SPP problem. Only the Vicuna-13b model showed slightly better performance, although not significant, on the EDP problem compared to SAS across all investigated models.

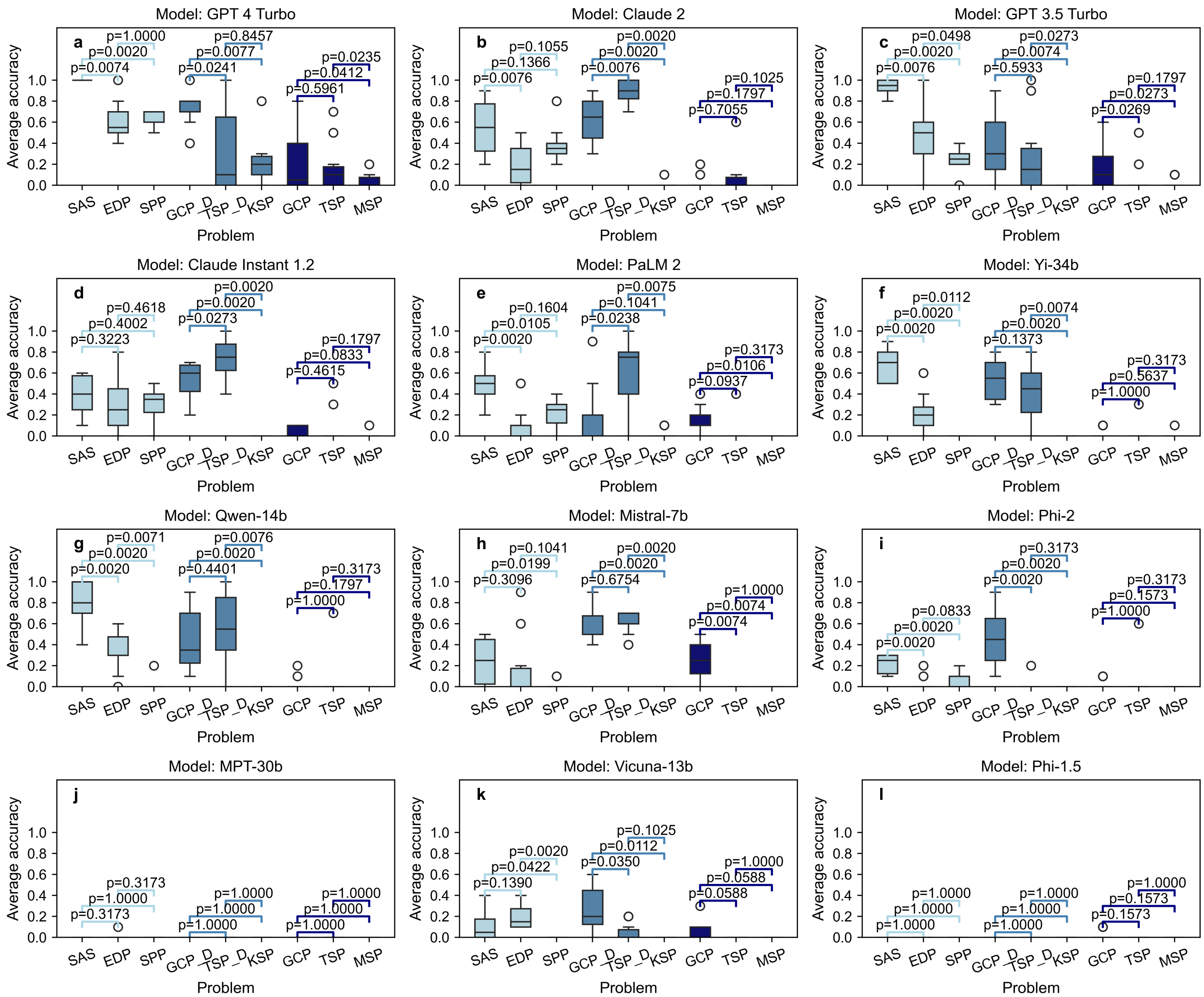

Figure 5.5: Models' performance on tasks across complexity levels. (a) GPT 4 Turbo. (b) Claude 2. (c) GPT 3.5 Turbo. (d) Claude Instant 1.2. (e) PaLM 2. (f) Yi-34b. (g) Qwen-14b. (h) Mistral-7b. (i) Phi-2. (j) MPT-30b. (k) Vicuna-13b. (l) Phi-1.5.

Other observations

GPT 4 Turbo showcased a very similar performance between the EDP and SPP problems, while Claude Instant 1.2 exhibited a similar performance for all three of these problems. Yi-34b, Qwen-14b, GPT 3.5 Turbo, and GPT 4 Turbo displayed remarkably high accuracy specifically for the SAS task. MPT-30b and Phi-1.5 showed very limited performance in identifying these three problems.

With regard to NP-Complete complexity, there are several observations to highlight. Still, neither MPT-30b nor Phi-1.5 could provide any identification of problems in the NP-Complete complexity. In the case of the GCP-D and TSP-D problems, the performance of these models varied significantly. Phi-2, Vicuna-13b, and GPT 4 Turbo outperformed in the GCP-D problem compared to TSP-D, whereas Claude Instant 1.2, Claude 2, and PaLM 2 exhibited better performance in TSP-D over GCP-D. However, models such as Mistral-7b, Yi-34b, Qwen-14b, and GPT 3.5 Turbo showcased relatively similar performance between these two tasks.

For the KSP task, only GPT 4 Turbo demonstrated promising performance, while the remaining models faltered.

Considering NP-Hard complexity as the most intricate task set among the three (as evidenced in Figure 5.4), many of the models examined encountered challenges in identifying tasks within this complexity. For the GCP task, Mistral-7b, PaLM 2, GPT 3.5 Turbo, and GPT 4 Turbo exhibited some potential, while Vicuna-13b and Claude Instant 1.2 showed limited performance. For the TSP task, identification was observed only in Claude 2 and GPT 4 Turbo. Of all the investigated models, GPT 4 Turbo exhibited promise in identifying these three tasks within the NP-Hard complexity. However, the performance in GCP and TSP identification significantly surpassed that of the MSP task across these models. For the MSP task, only GPT 4 Turbo displayed some ability for identification, while with notably low accuracy.

5.3 Learning Ability

In the exploration of machine learning, particularly with LLMs, distinguishing genuine learning from mere mimicry is critical, especially in complex problem-solving contexts. This section evaluates LLMs' in-context learning capabilities to determine if they genuinely acquire and apply algorithmic skills across varying difficulty levels within tasks, rather than just replicating problem-solving processes. Additionally, to enhance these capabilities, we fine-tune three open-source LLMs, Phi-2, Mistral-7b, and Qwen-14b, using varied benchmarks aimed at exceeding performance of top-tier closed-source models such as GPT 4 Turbo. This approach tests both adaptability and depth of learning, marking a significant step in the use of LLMs for advanced cognitive tasks.

5.3.1 In-context Learning

Given examples in the context, can LLMs genuinely learn and apply algorithmic skills presented in contextual examples as opposed to simply mimicking problem solving processes [304], [305]? We differentiate between "learning" and "mimicking" by evaluating whether LLMs can generalize solutions to new problems of varying difficulty levels within the same task, after being exposed to examples. Our hypothesis is that if an LLM has truly learned the underlying ability in a complex scenario (algorithmic skill for the benchmark), it should be able to tackle problems across different difficulty levels within the same task. Conversely, if an LLM is simply mimicking, its performance may falter when faced with variations in problem difficulty.

Experiment: Comparative analysis of learningability by in-context learning

A prevalent approach in current few-shot learning involves using examples that are similar to the test question. However, this raises a question about the extent to which the model is replicating the problem solving process from the examples, as opposed to genuinely acquiring reasoning skills. Consequently, it becomes pertinent to investigate whether the problem-solving abilities developed through example-based learning are generalizable.

To dig deeper into the models' in-context learning abilities, we utilize various few-shot in-context learning prompts to discern whether the model is "learning" from the few-shot examples or merely "mimicking" the behavior. In our benchmark, since we distinctly classify the difficulty level of each question, it allows for the use of questions from the same task but with varying difficulty levels as few-shot examples. The crux of this analysis lies in varying the difficulty levels of the examples within the prompts. Since the fundamental algorithmic skill required to solve a question remains constant across varying difficulty levels under the same task, a model that truly learns this skill should show consistent performance regardless of the example difficulty in the prompt. We propose the following hypotheses about the relationship between in-context learning ability and the difference of difficulty level between the given examples and the question being asked in context:

- Models possessing optimal generalization capabilities should demonstrate consistent performance improvement regardless of the difficulty level of the prompt examples in context. This assumption is based on the premise that models with robust learning abilities are capable of discerning and applying the intrinsic problem-solving skills learned in the examples. Given that questions within the same task fundamentally require similar skills, variations in difficulty are unlikely to significantly affect the model's performance.
- If a model exhibits the ability to generalize only from some types of examples but is unable to extend this learning to others, it reveals a deficiency in its capacity for generalization in terms of reasoning. This suggests that the model is not genuinely acquiring problem-solving skills from the examples but merely recognizing and applying patterns from examples that are of equal or greater complexity to the problem at hand.
- If a model is unable to generalize from either more difficult or easier examples and is restricted to examples of the same difficulty level, it strongly suggests that the model is merely replicating the process presented in the context rather than internalizing any fundamental problem-solving techniques or pattern recognition embedded within the examples. This behavior indicates a profound deficiency in the model's ability to comprehend and understand the underlying principles. It points to an absence of transferable, logic-learning skills, reflecting a superficial form of learning that is limited to surface-level imitation rather than a deeper, conceptual grasp.

We categorize the few-shot prompts into three types.

- Few-shot prompts with examples of the same difficulty level: Here, the model is provided with five examples in the prompt, all of which are at the same difficulty level and distinct from the question being asked.
- Few-shot prompts with examples that are easier than the question: This set comprises five variations of prompts, each with examples that are 1, 2, 3, 4, and 5 levels easier than the question, respectively.
- Few-shot prompts with examples that are more challenging than the question: Similarly, we prepare five sets of prompts, each containing examples that are 1, 2, 3, 4, and 5 levels more difficult than the question, offering a gradient of increased challenge.

Through this diverse array of prompts, we aim to provide a nuanced understanding of the LLMs' ability to learn from examples, thereby offering valuable insights into their underlying learning capabilities.

Results: Effects of a few-shot example difficulty on improving reasoning ability

Figure 5.6 illustrates the results of few-shot learning experiments in SAS and EDP in various models, each subjected to 11 distinct few-shot prompts. These prompts were systematically varied in terms of difficulty level, ranging from -5 to +5 relative to the difficulty level of the target question each model was tasked with solving. The results underscore a notable improvement in model performance while exhibiting significant variability across the different prompts on open-source models. Such observations indicate a potential limitation in the open-source models' capacity to acquire the underlying task-solving skills and to generalize the examples from prompts. This highlights a crucial area for further investigation and development.

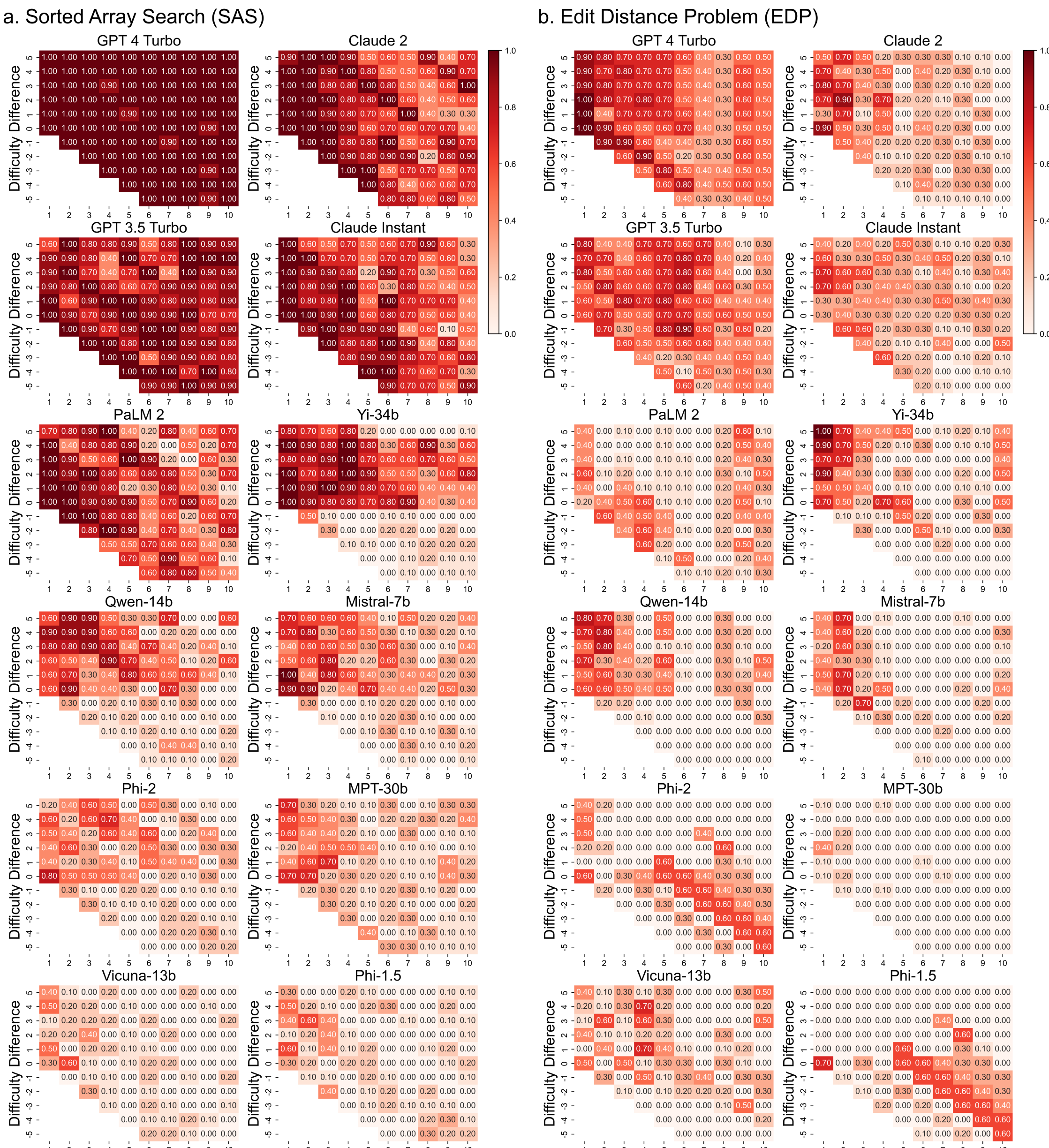

Figure 5.6: Few-shot learning results on SAS and EDP of all models

For closed source models, including GPT 4 Turbo, Claude 2, GPT 3.5 Turbo, PaLM 2, and Claude Instant 1.2, the results were notably close to the ideal scenario. We observed minimal variation in performance across different levels of difficulty in the examples provided. This consistency suggests that these models are not merely mimicking the solutions, but are indeed learning the algorithmic skills presented in the context of the examples.

In contrast, the performance of open-source models, particularly Yi-34b and Mistral-7b, exhibits a clear pattern where the models generally generalize well from examples that are more challenging than the given question, yet they struggle to do so from simpler examples. Other open-source models display less distinct patterns, but a notable trend is still evident: These models demonstrate some capacity to generalize from more challenging to simpler questions, but they are less successful in generalizing from simpler to more complex questions. An exception is observed with the Phi-1.5 model in EDP, where it appears to generalize better from easier examples than from harder examples at certain difficulty levels. However, broadly speaking, none of the open-source models consistently learns from both harder and easier examples. The difficulty level significantly influences the models' performance, suggesting a tendency for these models to mimic patterns rather than engage in genuine learning from the context. We document more details of this comparison between close- and open-source models in Tables 5.1 and 5.2. The overall results suggest that the open source models, especially the GPT and Claude family ones, are the suitable model candidate for reasoning in complex scenarios.

Accuracy	**GPT 4 Turbo**	**Claude 2**	**GPT 3.5 Turbo**	**Claude Instant**	**PaLM 2**	**Yi-34b**	**Qwen-14b**	**Mistral-7b**	**Phi-2**	**MPT-30b**	**Vicuna-13b**	**Phi-1.5**
Prompts on SAS												
Zeroshot	**1.000**	0.445	**0.942**	0.442	0.416	0.620	**0.706**	0.149	0.191	0.000	**0.113**	0.000
Fewshot (-5)	0.978	0.685	0.920	0.735	0.603	0.065	0.103	0.043	0.095	0.165	0.085	**0.155**
Fewshot (-4)	**1.000**	0.662	0.902	0.667	0.516	0.093	0.189	0.129	0.149	0.156	0.084	0.118
Fewshot (-3)	0.982	0.694	0.831	**0.769**	0.496	0.143	0.114	0.153	0.116	0.131	0.067	0.084
Fewshot (-2)	**1.000**	**0.771**	0.910	0.710	0.617	0.102	0.070	0.094	0.073	0.117	0.037	0.087
Fewshot (-1)	0.987	0.770	0.896	0.589	0.607	0.048	0.126	0.085	0.107	0.157	0.087	0.057
Fewshot (0)	0.984	0.671	0.846	0.651	**0.660**	0.598	0.255	**0.413**	0.222	0.258	0.089	0.098
Fewshot (1)	0.991	0.580	0.878	0.696	0.455	0.593	0.455	0.386	**0.287**	0.233	0.055	0.109
Fewshot (2)	**1.000**	0.675	0.829	0.587	0.656	0.647	0.444	0.296	0.260	0.175	0.067	0.056
Fewshot (3)	0.993	0.736	0.800	0.598	0.489	**0.662**	0.427	0.318	0.275	0.144	0.093	0.098
Fewshot (4)	**1.000**	0.729	0.869	0.580	0.471	0.638	0.251	0.287	0.195	**0.269**	0.053	0.106
Fewshot (5)	**1.000**	0.671	0.844	0.602	0.607	0.167	0.387	0.356	0.196	0.202	0.055	0.064
Prompts on EDP												
Zeroshot	0.536	0.120	0.318	0.176	0.033	0.166	**0.269**	0.058	0.009	0.002	0.147	0.000
Fewshot (-5)	0.387	0.075	0.417	0.048	0.170	0.000	0.000	0.015	0.210	0.000	0.000	0.205
Fewshot (-4)	**0.556**	**0.209**	0.367	0.102	0.207	0.000	0.000	0.000	0.300	0.000	0.044	0.284
Fewshot (-3)	0.500	0.178	0.386	0.167	0.235	0.029	0.000	0.029	0.327	0.000	0.108	0.331
Fewshot (-2)	0.462	0.173	0.479	0.210	0.208	0.146	0.065	0.090	0.329	0.000	0.154	**0.335**
Fewshot (-1)	0.485	0.200	0.513	0.246	**0.289**	0.135	0.069	0.098	**0.348**	0.011	**0.248**	0.328
Fewshot (0)	0.518	**0.209**	**0.564**	0.253	0.238	**0.282**	0.227	**0.182**	0.320	**0.022**	0.164	0.293
Fewshot (1)	0.535	0.184	0.535	**0.355**	0.205	0.089	0.266	0.089	0.115	0.013	0.160	0.115
Fewshot (2)	0.545	**0.209**	0.544	0.238	0.196	0.195	0.266	0.042	0.098	0.015	0.093	0.087
Fewshot (3)	0.536	0.189	0.449	0.315	0.182	0.127	0.140	0.067	0.060	0.007	0.191	0.051
Fewshot (4)	0.538	0.209	0.507	0.305	0.200	0.247	0.186	0.095	0.009	0.000	0.129	0.000
Fewshot (5)	0.531	0.205	0.449	0.244	0.167	0.271	0.146	0.055	0.015	0.009	0.202	0.000

Table 5.1: Weighted accuracy of Zero-shot and Few-shot on SAS and EDP. The best performance for each column is highlighted with bold font (respectively for SAS and EDP).

Failure Rate	GPT 4 Turbo	Claude 2	GPT 3.5 Turbo	Claude Instant	PaLM 2	Yi-34b	Qwen-14b	Mistral-7b	Phi-2	MPT-30b	Vicuna-13b	Phi-1.5
Prompts on SAS												
Zeroshot	0.000	0.260	0.000	0.400	0.110	0.330	0.200	0.070	0.150	1.000	0.480	1.000
Fewshot (-5)	0.000	0.060	0.020	0.000	0.000	0.940	0.900	0.480	0.920	0.160	0.640	0.860
Fewshot (-4)	0.000	0.033	0.000	0.000	0.017	0.917	0.817	0.617	0.867	0.117	0.633	0.900
Fewshot (-3)	0.000	0.057	0.057	0.014	0.000	0.871	0.886	0.471	0.886	0.043	0.600	0.914
Fewshot (-2)	0.000	0.075	0.025	0.000	0.013	0.888	0.913	0.538	0.900	0.063	0.613	0.888
Fewshot (-1)	0.000	0.044	0.022	0.000	0.011	0.911	0.856	0.622	0.878	0.078	0.589	0.933
Fewshot (0)	0.000	0.060	0.020	0.000	0.020	0.300	0.640	0.380	0.670	0.040	0.540	0.880
Fewshot (1)	0.000	0.040	0.020	0.010	0.010	0.290	0.500	0.420	0.700	0.060	0.520	0.830
Fewshot (2)	0.000	0.040	0.040	0.000	0.010	0.290	0.510	0.440	0.710	0.030	0.640	0.910
Fewshot (3)	0.000	0.020	0.050	0.010	0.030	0.280	0.450	0.460	0.670	0.030	0.650	0.830
Fewshot (4)	0.000	0.050	0.030	0.000	0.040	0.280	0.570	0.530	0.700	0.080	0.630	0.850
Fewshot (5)	0.000	0.050	0.040	0.000	0.020	0.680	0.520	0.440	0.740	0.050	0.650	0.900
Prompts on EDP												
Zeroshot	0.000	0.000	0.000	0.000	0.440	0.000	0.000	0.040	0.000	0.960	0.160	0.950
Fewshot (-5)	0.000	0.000	0.000	0.140	0.160	0.000	0.320	0.140	0.540	0.880	0.460	0.640
Fewshot (-4)	0.000	0.000	0.000	0.100	0.150	0.000	0.233	0.050	0.400	0.817	0.383	0.483
Fewshot (-3)	0.000	0.043	0.000	0.057	0.100	0.000	0.100	0.029	0.300	0.871	0.329	0.400
Fewshot (-2)	0.000	0.038	0.000	0.025	0.075	0.000	0.038	0.025	0.263	0.700	0.238	0.350
Fewshot (-1)	0.000	0.011	0.000	0.056	0.033	0.000	0.022	0.000	0.167	0.667	0.200	0.256
Fewshot (0)	0.000	0.020	0.000	0.060	0.000	0.000	0.040	0.000	0.130	0.730	0.190	0.200
Fewshot (1)	0.000	0.000	0.000	0.070	0.000	0.000	0.010	0.000	0.100	0.800	0.190	0.210
Fewshot (2)	0.000	0.040	0.000	0.050	0.000	0.000	0.020	0.000	0.000	0.750	0.090	0.100
Fewshot (3)	0.000	0.020	0.000	0.060	0.040	0.000	0.000	0.000	0.010	0.710	0.040	0.100
Fewshot (4)	0.000	0.030	0.000	0.030	0.000	0.000	0.000	0.000	0.000	0.810	0.000	0.000
Fewshot (5)	0.000	0.050	0.000	0.030	0.000	0.000	0.000	0.000	0.000	0.820	0.030	0.000

Table 5.2: Weighted failure rate of Zero-shot and Few-shot on SAS and EDP.

5.3.2 Finetuning

To enhance LLMs' learning capabilities, finetuning is another option for customize and contextualize models to a specific problem. In the finetuning experiment, the central hypothesis is that by dynamically updating these benchmarks, the models can achieve better performance than the best close source model GPT 4 Turbo. To test this, we finetuned three open-source LLMs—Phi-2, Mistral-7b, and Qwen-14b—on five different versions of the benchmarks. We then assessed the performance of each model's checkpoint on two benchmark versions that varied in complexity. This method helped determine if finetuning equips LLMs to effectively tackle benchmarks with different levels of difficulty.

Experiment: Finetuning Open Source Models

The experiment involved finetuning three high-performing open-source models: Phi-2, Mistral-7b, and Qwen-14b. Due to constraints in computing resources, the Yi-34b model was not included in the finetuning process. For the finetuning process, we employed the QLoRA technique, applying specific hyperparameters: batch size set to 8, a single epoch, a warmup proportion of 0.03, a learning rate of 1e-4, lora_r at 64, lora_alpha at 16, and a lora_dropout of 0.1.

Results: Effects of Model Finetuning on Reasoning Ability Enhancement

Figure 5.7: Model's robustness on different problems and difficulty levels.

We explore the model learning ability through a process of finetuning on pairs of question and gold answer. We experiment using 3 well-performing open-source models: Qwen-14b, Mistral-7b, and Phi-2 on two versions of benchmarks. Figure 5.7 presents the result[2]: each problem has two graphs with one displaying evaluation results at difficulty levels 1-10 and one displaying evaluation results at difficulty levels 11-20. In each graph, the first row indicates the mean accuracy of each model, averaged over the outcomes at 5 finetuning checkpoints, ranging from tuning using zero (no finetuning) to five distinct benchmarks. Our findings are in twofolds as the follow.

First, while finetuning yields improvements in solving polynomial-time problems, its impact on the more complex NP-complete and NP-hard problems are negative. This suggests the inherent difficulty of hacking NP-complete, and potentially NP-hard, problems through the basic finetuning with question-and-answer approach. Manual annotation of the chain-of-thought, which is not provided in the benchmarks, could potentially enhance effectiveness, but with challenges in annotation.

[2]We do not present the result on MSP as this problem does not have a fixed solution and no finetuning was conduct on it.

Second, finetuning appears beneficial for performance within the same difficulty level of all P problems, yet shows limited out-of-distribution (OOD) adaptability and struggles to generalize to more difficult problems (as evidenced in graphs a and c) except for SAS. For instance, Qwen-14b demonstrates notable proficiency on SPP challenges at levels 1-10 following finetuning; its performance is comparable to that of GPT-4. However, its performance significantly diminishes on SPP problems at levels 11-20, even underperforming compared to its unfinetuned checkpoint. This indicates that finetuning on these benchmarks can only benefit very simple questions such as SAS but could potentially impede generalization capabilities and render finetuning hacking useless.

In conclusion, the current candidate open source LLMs are unable to significantly improve performance through finetuning due to two primary factors: (1) the inherent complexity of NP-complete and NP-hard problems, which are difficult to learn solely from question-answer pairs, and (2) the propensity for P problems to become overfitted through finetuning on these pairs, while the real "reasoning" ability can be easily exposed by increasing the problem difficulty level. Thus, for finetuning tasks in the GAUDS system, we choose the highly competent close source model, GPT 3.5 Turbo[3], as an alternative to the open source ones.

5.4 Choice of (Candidate) Models for the GAUDS System

In the development of advanced data search systems like GAUDS, the choice of underlying models plays a crucial role, especially when the system requires complex reasoning capabilities and the ability to understand and translate natural language into database queries effectively. For our specific tasks, which include topic classification and reuse generation, the performance of the LLMs used is paramount. These tasks not only demand a deep understanding of the language but also require the models to execute precise translations of user queries into a form that can be comprehended and acted upon by the system's database.

This chapter's results point to the models from OpenAI and Anthropic, specifically the GPT and Claude families, as being particularly well-suited for these tasks. These models have consistently demonstrated superior performance in complex reasoning scenarios compared to their counterparts. This performance edge is attributed to their advanced architectures, which are finely tuned to handle a broad range of language processing tasks with greater accuracy and efficiency. The models' ability to engage with and process detailed linguistic structures makes them ideal candidates for a system that aims to provide precise and contextually aware responses to user queries.

The need for high-performance LLMs in the GAUDS system is driven by the complexity of the tasks it aims to manage. Topic classification involves discerning subtle nuances in language to categorize content accurately, while reuse generation requires the model to understand deeply the context of existing data to suggest relevant uses for new research scenarios. These tasks are not only linguistically demanding but also require a level of semantic understanding that earlier models might struggle to provide.

Therefore, based on the empirical data and comparative analyses presented, we have chosen to incorporate the latest iterations of LLMs from the GPT and Claude families into our experiments. Each of these models brings a unique set of capabilities that are crucial for the success of the GAUDS system, particularly in

[3]GPT 4 Trubo is not a feasible candidate for finetuning due to the cost and tier request by OpenAI

handling the intricacies of language-based user interactions and ensuring the system's outputs are both relevant and accurate. The technical specifications of these models, as detailed in After our experiments in GPT and Claude models in late 2023, they also update the models with state-of-the-art data and abilities. The continuous maintaining and developing of these models affirm their applicability for our system, providing assurance that they can handle the tasks required by GAUDS efficiently and effectively even when tasks evolve in complexity.

By selecting these models, we ensure that the GAUDS system is equipped with the most advanced tools available for natural language processing, setting a strong foundation for the system's ability to deliver exceptional performance in data discovery and reuse tasks. This strategic choice not only aligns with our goal to enhance the system's capabilities but also ensures that as the system evolves, it remains at the cutting edge of technology, capable of adapting to new challenges and expanding for data user needs.

Chapter 6

Building the Generative AI-augmented and User-centric Data Search (GAUDS) System

In the quest to advance the utility and accessibility of scholarly research data, the development of the GAUDS system represents a pivotal evolution. This system builds on the foundational technologies of SKGs and customized LLMs to offer an innovative approach to data search and reuse. Drawing from the experiences and insights gained from the prototypes of the data search system (in Chapter 3), the GAUDS system has been meticulously designed to address and overcome the limitations previously encountered in academic data discovery environments.

The journey to GAUDS began with prototypes like SimSearch (see Section 3.2) and DataChat (see Section 3.3), which served as initial attempts to enhance data discovery through AI and NLP technologies. From SimSearch, valuable lessons were learned about the complexity of effectively mapping and retrieving data through simplistic semantic models, often restricted by the static and finite nature of resources such as WordNet [229]. DataChat expanded on these concepts by incorporating more dynamic LLMs to facilitate direct data-user interactions, providing insight into the importance of context and user-specific customization in data searches [290]. These prototypes highlight the challenges inherent in developing semantic agents that could adequately handle the diversity and depth of academic datasets, particularly when dealing with new and evolving topics. The lack of flexibility in traditional semantic similarity matching and the challenges due to the complexity in database schema prohibit the development of comprehensive and effective data search systems. These experiences underscored the need for a more robust and adaptable framework, capable of not only understanding but also anticipating the complex needs of users in a scholarly environment.

Informed by the inflexibility of semantic matching and the complexity of the database schema in data search prototypes, GAUDS is designed to seamlessly integrate SKGs and LLMs to create a more intuitive and effective search experience. The system is rooted deeply in the literature on data structures and user-centric design principles (Chapter 3), ensuring that each feature is aligned with the real-world needs of researchers. Each design decision within GAUDS—from the use of advanced LLMs for query processing to the implementation of SKGs for managing relational data—is aimed at maximizing the system's relevance and utility in an academic context.

GAUDS addresses previously identified challenges by implementing enhanced NLP tools and more comprehensive data integration techniques. For example, moving beyond the limitations of WordNet, GAUDS employs more sophisticated semantic models such as BERT and transformer-based approaches, which provide greater accuracy and flexibility in understanding and processing complex academic queries [306], [307]. Furthermore, the system has been designed to better accommodate the diversity of scholarly data and metadata, addressing the inclusiveness and diversity issues previously noted in semantic mapping (see Section 3.2).

As we delve deeper into the architecture and functionality of the GAUDS system in this chapter, the subsequent sections will first overview the system, followed by detailed discussions including the design decisions, the underlying models and heuristics, and the intelligent nature of the system with regard to its key features. These features include exploratory discovery through topic classification, directed discovery through SKG-based question answering, data recommendation with vector search, and guided discovery pathways for data reuse, all of which are aimed at enhancing connectivity, effectiveness, visibility, and interactivity. Finally, we will study cases with specific health and medical research questions to exemplify the GAUDS system's ability to support research data discovery and reuse.

6.1 Overview of the GAUDS System

We provide an overview of the GAUDS system from both the system and the user perspectives.

6.1.1 System Workflow

The GAUDS system operates as a cohesive unit, incorporating both exploratory and directed discovery mechanisms along with advanced retrieval-augmented generation techniques for data recommendation. Figure 6.1 illustrates the overall workflow of the GAUDS system, highlighting the integration of user inputs, backend processing through LLMs, and the SKG to deliver precise and contextually relevant search results.

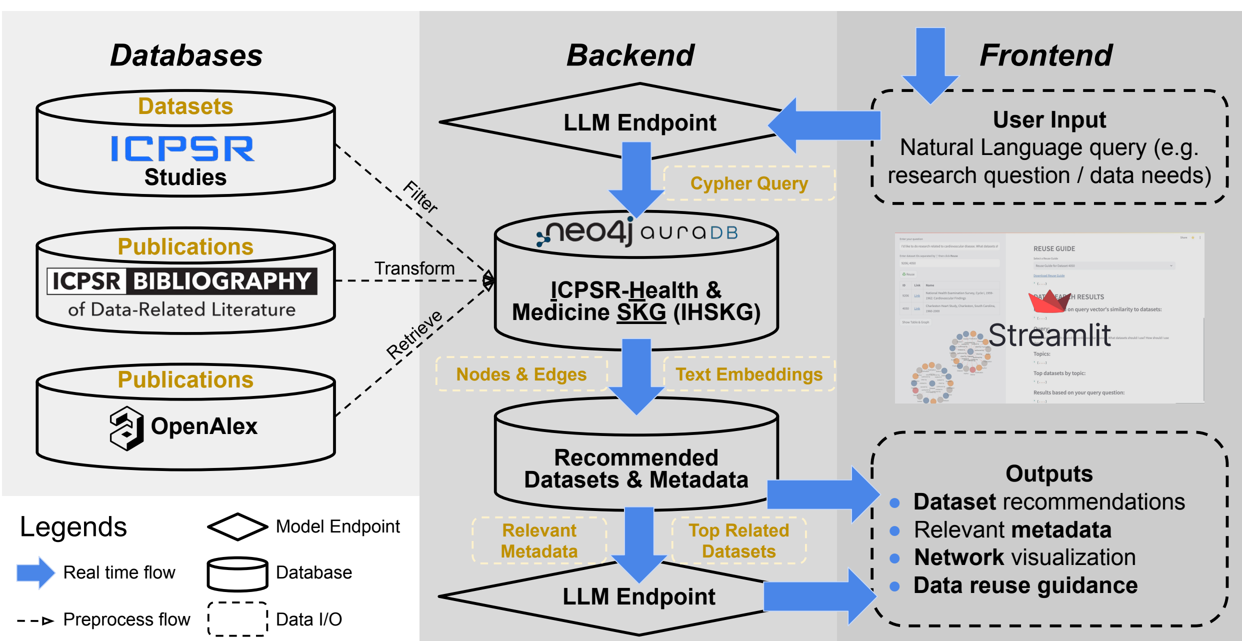

Figure 6.1: GAUDS System Workflow

The workflow begins with user inputs in the form of natural-language queries, which may include research questions or specific data needs. These inputs are processed through an LLM endpoint, which interprets the queries and generates outputs that include recommendations of the dataset, relevant metadata, network visualizations, and data reuse guidelines. To ensure the relevance and precision of the recommendations, the system employs Cypher queries that retrieve data from IHSKG and filter the outputs based on the underlying schema and conditions.

The backend of the system consists of ICPSR studies and bibliographies databases that store datasets and publications, which are crucial for the retrieval process. These data are transformed and retrieved in a static manner, allowing the system to generate text embeddings, as we introduced in Chapter 4, in an economic way. These text embeddings facilitate the matching of user queries with the most relevant datasets and metadata based on semantic cosine similarity [308][1]. In addition, the integration of OpenAlex [264] within the system provides an additional layer of data, enhancing the system's ability to recommend related publications and authors and to extend its knowledge base.

Overall, the GAUDS workflow is designed to maximize efficiency and accuracy in data retrieval and recommendation, aligning with the system's goal of providing researchers with intuitive and effective tools for discovering and reusing scholarly data.

6.1.2 User Interface and Paths

Figure 6.2 shows the design of the GAUDS system, which includes two primary areas: the sidebar, on the left with a gray background designed for user input, and the main interface, on the right with a white background where the results are displayed. The interface also includes a how-to guide that provides users with step-by-step instructions on how to utilize the system for dataset discovery and reuse.

[1]We choose the cosine similarity because it works well in vector-based semantic comparison, which is different from path or similar metrics for WordNet (introduced in Section 3.2). Cosine similarity also works well with the sparse vectors generated by OpenAI's 'text-embedding-3-small' embedding model (discussed in Chapter 4)

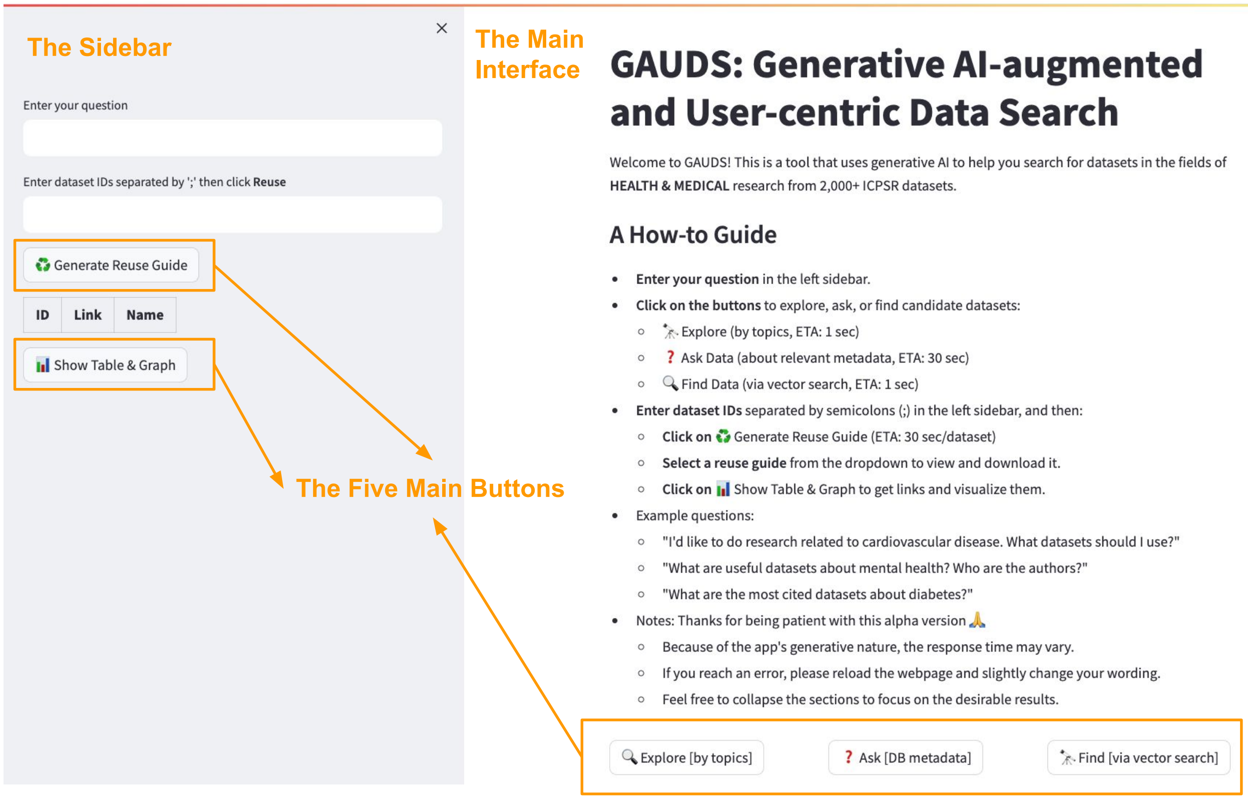

Figure 6.2: GAUDS System Frontend

The interface is navigated via five main buttons: **Explore** (by topics), **Ask** (DB metadata), **Find** (via vector search), **Generate Reuse Guide**, and **Show Table & Graph**. These buttons facilitate a seamless translation of natural language queries into customized dataset discovery and reuse experiences, guiding the user from initial inquiry to detailed data exploration.

The GAUDS system serves discovery and reuse functionalities, as illustrated in Figure 6.3. After users engage with the Explore, Ask, or Find functionalities to gather dataset-related information, the system provides relevant data and metadata outputs. Users can interactively refine their natural language queries if they are not satisfied with the recommendations results. When satisfied with the recommended datasets, users can then input datasets of interest into the Reuse Guide Generator, which provides comprehensive details including Fit for Purpose assessments (Section 6.5.1) and Data Documentation (Section 6.5.2), and utilize the Show Table & Graph feature to visualize the data (Section 6.5.3).

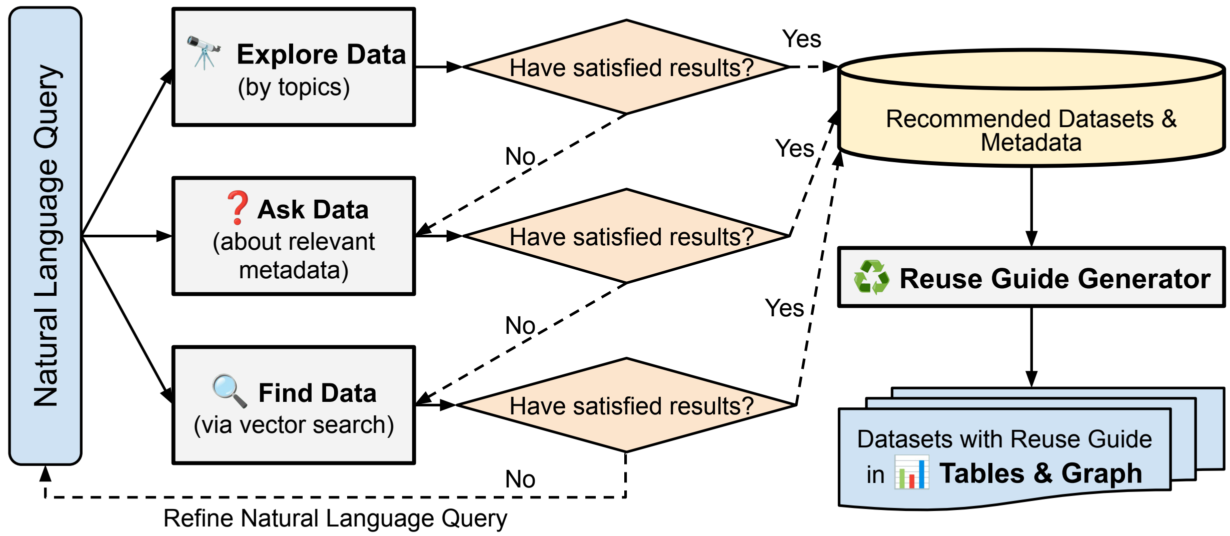

Figure 6.3: GAUDS User Paths

This structured approach not only enhances user interaction through clear navigational paths, but also supports varied user needs—from casual exploration of topics to detailed technical inquiries about specific datasets. By accommodating different types of user interaction, the GAUDS system ensures that users can locate and utilize datasets in a manner that best suits their research purposes. Moreover, the integration of these features allows users to effortlessly transition between different modes of data discovery and reuse, optimizing their research workflow, and enhancing productivity.

6.2 Explore Data: Exploratory Discovery through Topic Classification

The first step in building GAUDS, based on the IHSKG, is to customize LLMs for classifying a natural language query to a research topic. The research topic can help to map datasets with the same topic to the user query.

6.2.1 Data and Methods

In this section, we discuss the data sampling framework and methods to improve and evaluate the performance of LLMs in the discovery of exploratory research data.

Models and Learning Methods

For the development of the GAUDS system, we start with five LLMs as candidates and use customization methods, including zeroshot learning, fewshot learning, and finetuning, to improve the performance with respect to data discovery methods.

Models from OpenAI and Anthropic, in particular, the GPT and Claude family models, are outperforming most of other closed and open source LLMs in complex reasoning tasks, as we discussed in Chapter 5. For our tasks of topic classification (and reuse generation in the later section), there is a strong need for high performance LLMs for understanding the natural language and translating it to the database language. Thus, we choose the latest five LLMs in the GPT and Claude families for our experiment. Table D.1 shows their technical specifications [2], which indicates suitability for contextualized and economical implementation in the GAUDS system, where the user queries are often represented in tokens much fewer than the maximum context token limits and each search action may only cost below $0.0001.

Table 6.1: Specifications of LLMs

Name	**Platform**	**Version Date**	**Input Cost / M tokens**	**Output Cost / M tokens**	**Context Window**
GPT-4-Turbo	OpenAI	2024-04-09	$10	$30	200K
GPT-3.5-Turbo	OpenAI	2024-01-25	$5	$15	16K
Claude-3-Opus	Anthropic	2024-02-29	$15	$75	200K
Claude-3-Sonnet	Anthropic	2024-02-29	$3	$15	200K
Claude-3-Haiku	Anthropic	2024-03-07	$0.25	$1.25	200K

Zeroshot learning is the default setting in our experiments of topic classification, where we provide a prompt of agent role (data reuse consultant), the steps before generating the output, the output format, and the natural language query to an LLM. We require the LLM to output in XML format for accurate result parsing, which was empirically proved to be effective in Chapter 5. Below is the overall structure of our prompt (for simplicity, we abbreviated the list of 68 topics and the natural language query in the prompt structure):

```
You are a data reuse consultant. Please follow the thought process below to find all
    highly relevant topics for a project summary.

Step 1: Read the summary of a project (query text) at the end of the prompt.

Step 2: Find the highly relevant topics for the project summary from the list of topics
     below (seperated by ;):
...... [the topic list]

Note: Usually, there should be around 3 to 5 topics.

Output format: The results should be returned as a list of strings between XML tags <
    result> </result> and seperate topics by "; ". For example, <result>[Genetics;
    Endocrinology, Diabetes and Metabolism; Public Health, Environmental and
    Occupational Health]</result>

```

[2]The technical specifications are subject to change by the LLM API endpoint; the latest pricing information can be found on their websites: https://openai.com/pricing and https://www.anthropic.com/api which are likely lower than the presented cost

```
The summary of a project (query text) is:...... [the natural language query]
```

Listing 6.1: Topic Classification Zeroshot Prompt

Fewshot learning can further enhance the robustness and effectiveness of the topic classification task. This methodology involves providing the language model with a small number of examples before it performs the classification task, which helps in setting the context and refining its predictions based on the provided examples. In the fewshot setup, we expand on the zero-shot configuration by providing examples of project summaries along with their correctly classified topics. Each example acts as a guide, showing the model what kind of output is expected when it encounters a similar type of query. Below is a schematic representation of the fewshot prompt structure, where each example is detailed before the new query is presented:

```
Query Text: The association between demographic and behavioral characteristics and
    sunburn among U.S. adults - National Health Interview Survey, 2010
Topic Labels: ["Dermatology"]

Query Text: Do changes in sex steroid hormones precede or follow increases in body
    weight during the menopause transition? Results from the Study of Women's Health
    Across the Nation
Topic Labels: ["Endocrinology, Diabetes and Metabolism"]

Query Text: What newly licensed registered nurses have to say about their first
    experiences
Topic Labels: ["General Health Professions"; "Research and Theory"; "Emergency Medical
    Services"]

Query Text: Association between non-medical and prescriptive usage of opioids
Topic Labels: ["Pediatrics, Perinatology and Child Health"; "Anesthesiology and Pain
    Medicine"; "Public Health, Environmental and Occupational Health"]

Query Text: Unpacking self-rated health and quality of life in older adults and elderly
     in India: A structural equation modelling approach
Topic Labels: ["General Health Professions"]

Query Text: Facial fractures in young adults: A national retrospective study
Topic Labels: ["Ophthalmology"; "Surgery"]

Query Text: Mental disorders, gun ownership, and gun carrying among soldiers after
    leaving the Army, 2016-2019
Topic Labels: ["Occupational Therapy"]

Query Text: Temporal trend of cadmium exposure in the United States population suggests
     gender specificities
Topic Labels: ["Nutrition and Dietetics"]

Query Text: The brief window of time comprising a wheelchair transfer confers a
    significant fracture risk on elderly Americans
```

```
Topic Labels: ["Emergency Medicine"; "Public Health, Environmental and Occupational Health"]

Query Text: A rose by any other name? Objective knowledge, perceived knowledge, and adolescent male condom use
Topic Labels: ["General Health Professions"; "Infectious Diseases"]
```

Listing 6.2: Topic Classification Fewshot Examples

The use of ten examples in each group aims to sufficiently prime the model with a diverse set of scenarios, improving its ability to generalize from these examples to new, unseen project summaries. The hypothesis is that by exposing the model to several targeted examples, it will learn to more accurately discern the nuances between different topics and better align its outputs with the expected results. To further avoid randomness, we experiment for five groups, all using different examples as the above but in the same format.

Finetuning is a another post-customization method that enhances model performances. We adopt a methodologically sound approach, leveraging publication names as queries with the corresponding cited datasets serving as the ground truth for outcomes. This approach is reflected in the structure of the training data, which is constructed to mirror the specific needs outlined in OpenAI's fine-tuning data structure requirements [309]. In total, we used 386 randomly selected training samples based on the stratified sampling frame (will introduce in Section 6.2.1). Each training example is structured to include a publication title as the query and the relevant research topics as the expected output, serving as the true results for model training. Here, we provide three examples of finetuning samples:

```
# The first example
{"messages": [{"role": "system", "content": "You are a data reuse consultant. You need to predict research topics, based on a summary of project (query text)."}, {"role": "user", "content": "Prevalence of alopecia areata in the First National Health and Nutrition Examination Survey"}, {"role": "assistant", "content": "Dermatology; Urology"}]}

# The second example
{"messages": [{"role": "system", "content": "You are a data reuse consultant. You need to predict research topics, based on a summary of project (query text)."}, {"role": "user", "content": "A novel cutoff for the waist-to-height ratio predicting metabolic syndrome in young American adults"}, {"role": "assistant", "content": "Cardiology and Cardiovascular Medicine; Public Health, Environmental and Occupational Health"}]}

# The third example
{"messages": [{"role": "system", "content": "You are a data reuse consultant. You need to predict research topics, based on a summary of project (query text)."}, {"role": "user", "content": "A comparison of two measures of quality of life: their sensitivity and validity for patients with advanced cancer"}, {"role": "assistant", "content": "Surgery; Oncology; Public Health, Environmental and Occupational Health"}]}

```

```
# The fourth example
{"messages": [{"role": "system", "content": "You are a data reuse consultant. You need
    to predict research topics, based on a summary of project (query text)."}, {"role":
     "user", "content": "Socioeconomic disparity in adult mortality in India:
    Estimations using the orphanhood method"}, {"role": "assistant", "content": "
    Pediatrics, Perinatology and Child Health; General Health Professions"}]}
```

Listing 6.3: Training Data Examples for Topic Classification Finetuning

The first example shows the publication and topic pair between ”Prevalence of alopecia areata in the First National Health and Nutrition Examination Survey” and clinical-focused topics including ”Dermatology” and ”Urology”. The model is supposed to make a correct prediction for the highly professional and unique range of topics in health and medical domains.

The second example use the publication with ”A novel cutoff for the waist-to-height ratio predicting metabolic syndrome in young American adults.” The model is expected to accurately classify this publication into pertinent topics like ”Cardiology and Cardiovascular Medicine” and ”Public Health, Environmental and Occupational Health.” Similarly, the third example features the title ”A comparison of two measures of quality of life: their sensitivity and validity for patients with advanced cancer,” which is associated with topics in ”Surgery,” ”Oncology,” and ”Public Health, Environmental and Occupational Health.” Here, the challenge lies in the model’s ability to discern and categorize publications that bridge clinical practices and broader health impacts, highlighting the need for a model that can navigate interdisciplinary nuances effectively. Both examples underscore the importance of precise topic identification from publication titles that may encompass both the social science and the clinical ends in health and medical domains.

The fourth example, ”Socioeconomic disparity in adult mortality in India: Estimations using the orphanhood method,” is categorized under ”Pediatrics, Perinatology and Child Health” and ”General Health Professions.” This instance illustrates the model’s task to identify and classify research that combines demographic studies with health outcomes, which requires understanding the topics behind both the geographic and professional context of the research, demonstrating the coverage of additional topical complexity in IHSKG.

Data Sampling Frame

The Testing Set is created with a stratified sampling frame among all ICPSR publications and their metadata. We apply this universal testing set to all the topic classification experiments. An integral component of the dataset is the topics, which classifies each publication under a primary topic among the three top topics as we introduced in Chapter 4. For each topic listed, the gold dataset is filtered to a topical subset that keeps publications corresponding to the current topic of iteration. Figure 6.4 shows the overall distribution of the number of publications by health and medical topics, which has a long tail and follows the Lotka’s frequency distribution of scientific productivity [310], thus requiring sampling to balance the training set.

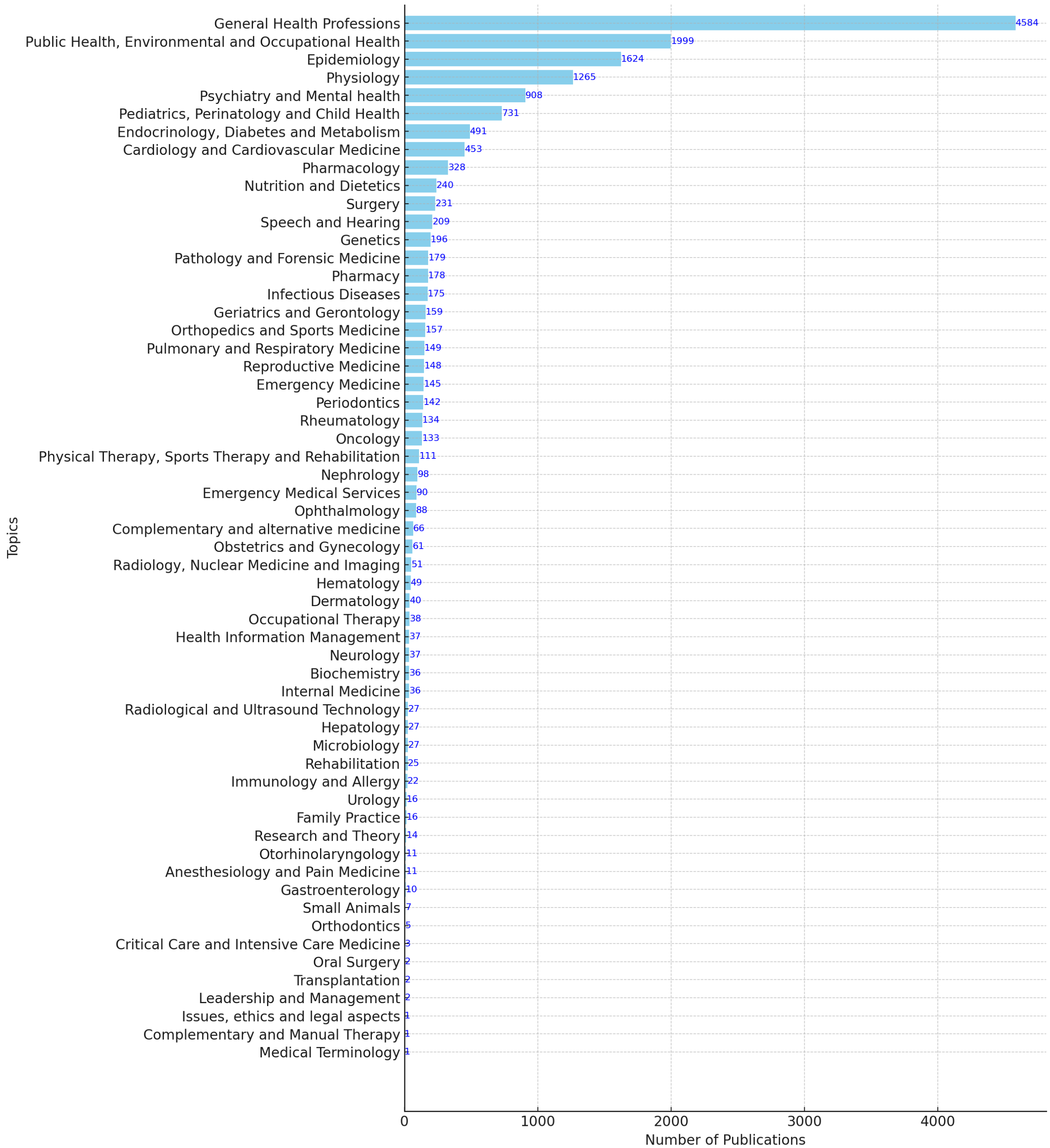

Figure 6.4: Number of Publications by Health and Medical Topics in IHSKG

Depending on the size of this subset, a conditional sampling method is employed. If the subset encompasses four or fewer publications, all entries are directly transferred to the sample. Conversely, if the subset contains more than 4 publications, a random sample of four publications is selected using the sample method, with a random state of 42 to ensure reproducibility of the sampling process across different executions. This sampling process is repeated across all topics, thereby incrementally assembling the final testing set. Upon completion of the iteration over all topics, we obtained a total of 216 data points, which are also used for the experiments in the next sections.

This sampling methodology ensures each topic is represented by up to four publications in the resultant samples. Where a topic's publication count does not exceed four, the complete set is included, ensuring comprehensive representation. For topics with more than four publications, the random sampling mitigates potential overrepresentation, thereby maintaining an equitable balance across the dataset. By fixing the random seed, the method assures that the sampling remains consistent and replicable. However, this approach may introduce selection bias by potentially omitting relevant publications from topics that are more extensively documented. For the training set, we avoid those topics with no more than four samples, to avoid data leakage in evaluation. This sampling strategy is particularly useful for analyses where uniform representation across topics is requisite, supported by IHSKG's summary statistics in Chapter 4, effectively preventing dominance by voluminous topics in the analytical outcomes. This balanced representation is crucial for conducting analyses that require equal input from each topic category, without skewing results towards more frequently documented topics.

The Fewshot Examples and Finetuning Samples In the development of few-shot learning datasets for publication classification based on topics, another stratified sampling strategy was employed to ensure balanced representation across various topics. Each dataset iteration generated a fixed number of examples – specifically two publications per topic, unless fewer were available. This procedure was replicated five times, each yielding a unique dataset to facilitate rigorous testing and validation of few-shot learning models. The random sampling was controlled via incrementally adjusted seeds to maintain reproducibility while introducing variability in the samples.

We apply a similar sampling frame for the finetuning data, while we choose six datasets per topic but remove the overlapping samples with the testing set. This sampling mechanism results in 386 training samples. Such an approach not only guarantees equitable topic representation but also enhances the robustness and generalizability of the learning models by providing diverse yet consistent training conditions.

These stratified sampling frameworks are particularly crucial for preliminary assessments in few-shot learning and finetuning scenarios, where limited amount of gold data often pose significant challenges.

Evaluation Metrics

In this section, we discuss the evaluation metrics for assessing the performance of the topic classification models, essential for ensuring the models' accuracy in categorizing academic publications correctly.

Precision, Recall, and F1 Score Precision, recall, and the F1 score are fundamental metrics used in the evaluation of classification models, particularly in the context of information retrieval and machine learning. These metrics help in assessing the accuracy and robustness of the classification algorithms:

- **Precision** (P) quantifies the number of correct positive predictions made by the model, calculated as:

$$P = \frac{TP}{TP + FP}$$

  where TP represents true positives and FP denotes false positives.

- **Recall** (R), or sensitivity, measures the model's ability to identify all relevant instances within a dataset:

$$R = \frac{TP}{TP + FN}$$

  where FN stands for false negatives.

- **F1 Score** is the harmonic mean of precision and recall, providing a balance between the two by considering both false positives and false negatives:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

These metrics are particularly important in the domain of scholarly data retrieval, where precision can help ensure that only relevant topics are associated with a publication, recall ensures that all relevant topics are captured, and the F1 score balances these aspects.

Any-Label Match Accuracy (ALMA) ALMA (Any-Label Match Accuracy) is a label-based binary accuracy metric designed to assess the performance of multi-label classification systems. The metric evaluates each instance by checking for any overlap between the predicted labels and the true labels. If at least one label matches, the instance receives a score of 1; otherwise, it scores 0. The overall ALMA score is then computed as the average of these individual scores across all instances. It is calculated as follows:

$$ALMA = \frac{1}{N} \sum_{i=1}^{N} match(Y_i, \hat{Y}_i)$$

where N is the total number of instances, Y_i is the set of true labels for the i-th instance, $\hat{Y}_i$ is the set of predicted labels for the i-th instance, and $match$ is a function that returns 1 if there is an overlap between Y_i and $\hat{Y}_i$, and 0 otherwise.

ALMA is particularly useful for evaluating systems where the presence of even a single correct prediction is valuable. This metric is developed to assist in the selection of related candidates of nodes in a knowledge graph for further investigation. Given that datasets in scholarly databases are often categorized under multiple topics, ALMA provides an optimistic estimation of the model's effectiveness in matching publications to relevant topics. This metric ensures that even partial correct predictions are acknowledged, making it highly suitable for complex multi-label classification tasks in dataset reuse settings.

6.2.2 Results

Category	Model	Precision	Recall	F1	ALMA
Zeroshot	GPT-4-Turbo	0.2	*0.41*	0.26	*0.69*
	Claude-3-Opus	0.23	0.35	0.27	0.62
	GPT-3.5-Turbo	0.24	0.3	0.25	0.54
	Claude-3-Sonnet	0.23	0.35	0.27	0.62
	Claude-3-Haiku	0.2	0.3	0.23	0.56
Fewshot with GPT-4-Turbo	Example Group 1	0.21	0.4	0.27	0.68
	Example Group 2	0.22	0.41	0.27	0.7
	Example Group 3	0.21	0.41	0.27	0.69
	Example Group 4	0.22	0.42	0.27	0.7
	Example Group 5	0.21	0.41	0.27	0.68
Finetuned GPT-3.5-Turbo	Fintuned Model	**0.47**	**0.56**	**0.49**	**0.82**

Table 6.2: Evaluation of Topic Classification Models and Learning Methods

Table 6.2 shows the results from the topic classification models, segmented into zeroshot learning, fewshot learning with GPT-4-Turbo, and finetuning with GPT-3.5-Turbo, reveal significant insights into the performance enhancements achievable through different LLM learning methodologies.

In the zeroshot learning setup, the performance metrics across various models such as GPT-4-Turbo, Claude-3-Opus, GPT-3.5-Turbo, Claude-3-Sonnet, and Claude-3-Haiku show a general trend of low effectiveness. The highest recall noted was 0.41 by GPT-4-Turbo, paired with a precision of 0.2, leading to an F1 score of 0.26 and an ALMA score of 0.69. The other models varied slightly in their precision and recall but generally hovered around similar low F1 scores, indicating modest capability in accurately classifying topics without prior examples or training specific to the task. The zeroshot results also shows that GPT-4-Turbo is the most promising model for further customization, since it outperforms the Claude-3-Opus model by nearly 20% in recall and more than 10% in ALMA. This observation confirms our selection of the GPT-4-Turbo for further fewshot learning.

Transitioning to fewshot learning with GPT-4-Turbo, where ten examples were provided in each group to prime the model, the results demonstrated slight improvements. All groups exhibited very similar outcomes, with F1 scores consistently at 0.27. Despite the introduction of example-based learning, the increments in performance metrics were marginal, suggesting that the limited number of examples was insufficient to significantly enhance the model's predictive accuracy. The ALMA scores ranged between 0.68 and 0.7, indicating a slight improvement in model confidence but not a substantial enhancement in performance.

The most noteworthy improvement was observed in the finetuned GPT-3.5-Turbo model, which markedly outperformed the zeroshot and fewshot setups. This model achieved a precision of 0.47 and a recall of

0.56, which culminated in an F1 score of 0.49—significantly higher than those observed in other models. Additionally, the ALMA score of 0.82 underscored a robust alignment with the human judgment in the gold dataset, affirming the efficacy of finetuning.

Figure 6.5 shows the finetuning process: based on the GPT-3.5-Turbo model, the training reached a 0.0306 training loss at the final step of 1158, while the loss decrease is fast – reducing from 4 to 1 in only 28 steps after the training starts, indicating a robust selection of training settings.

Figure 6.5: Finetuning Loss of GPT-3.5-Turbo

6.2.3 Discussion of the Explore Data Feature

As the results indicate, overall, while zeroshot learning provides a baseline capability in topic classification, the performance is considerably enhanced through finetuning. Fewshot learning, although intended to bridge the gap between zeroshot and finetuned scenarios, did not yield significant benefits under the conditions of this experiment. This underlines the possibility that more robust or numerous examples might be required for fewshot learning to effectively improve model performance in complex classification tasks, which diminishes the original purpose of the ”few” shot’s simplicity. Thus, for specialized tasks such as topic classification in fields like health and medical research, the investment in model training and customization appears to yield substantial benefits, enhancing both the precision and reliability of the models. We provide the following detailed analysis for each of the three learning methods.

We choose the finetuned GPT-3.5-Turbo model[3] as the final in-production model for the GAUDS system. Through the finetuning, it not only enhances the model’s ability to perform specific topic classification but also aids in generalizing across various academic fields, thus ensuring that the models are robust and effective in practical, scholarly search and discovery applications. This methodological rigor is essential for developing models that support advanced data retrieval tasks in scholarly communications, facilitating enhanced access to relevant research literature based on accurately identified topics.

6.3 Ask Data: Directed Discovery through SKG-based Question Answering

The second option for users to discover datasets is through directed discovery. While the directed search in relational databases is often directly related to a data or publication attribute, such as research funder, author, or geographic locations [56], directed discovery in IHSKG transcends these typical attributes by utilizing the

[3]The fintuned model is avaliable on OpenAI, named ”hemphill-research-group:lf:9CJL6uoL”

more interconnected and semantically rich graph structure. In this section, we develop the Ask Data feature of GAUDS, where we use LLMs to understand the schema of the SKG and identify relevant attributes within the IHSKG.

6.3.1 Prompt Design

This subsection introduces two types of prompts: 1) SKG schema prompt; 2) Data and Metadata Search prompt [4].

IHSKG schema prompt

To enable effective interaction with the IHSKG, prompts are designed to facilitate schema exploration and understanding. The prompt aims to generate dynamic queries based on the evolving structure of the SKG, allowing users to navigate through complex data landscapes intuitively.

```
# The Natural Language Part
This is the schema representation of the Neo4j database.

Node properties are the following:
{node_props}

Relationship properties are the following:
{rel_props}

Relationship point from source to target nodes
{rels}

Make sure to respect relationship types and directions

# The Cypher Query Part as Strings in Python
node_properties_query = """
    CALL apoc.meta.data()
    YIELD label, other, elementType, type, property
    WHERE NOT type = "RELATIONSHIP" AND elementType = "node"
    WITH label AS nodeLabels, collect(property) AS properties
    RETURN {labels: nodeLabels, properties: properties} AS output
"""

rel_properties_query = """
    CALL apoc.meta.data()
    YIELD label, other, elementType, type, property
    WHERE NOT type = "RELATIONSHIP" AND elementType = "relationship"
    WITH label AS nodeLabels, collect(property) AS properties
    RETURN {type: nodeLabels, properties: properties} AS output
"""
```

[4] There prompt designs are converted and contextualized from a Neo4j developer blog: https://neo4j.com/developer-blog/generating-cypher-queries-with-chatgpt-4-on-any-graph-schema/

```

rel_query = """
    CALL apoc.meta.data()
    YIELD label, other, elementType, type, property
    WHERE type = "RELATIONSHIP" AND elementType = "node"
    RETURN {source: label, relationship: property, target: other} AS output
"""
```

Listing 6.4: SKG schema prompt

Data and Metadata Search prompt

Utilizing the schema described above, this prompt aids users in crafting specific Cypher queries to extract detailed information from the graph, promoting efficient data retrieval and supporting nuanced research inquiries.

````
Task: Generate Cypher queries to query a Neo4j graph database based on the provided
    schema definition.

Instructions:
Use only the provided relationship types and properties.
Do not use any other relationship types or properties that are not provided.
If you cannot generate a Cypher statement based on the provided schema, explain the
    reason to the user.

Schema:
{schema}

Note 1: Do not include any explanations or apologies in your responses. \n
Note 2: Limit the number of output to 10. \n
Note 3: Output the query between triple backticks (```) to format it as code.

Example output:
```cypher
MATCH (p:PUBLICATION)-[:CITE]->(d:DATASET)
WITH d, COUNT(p) AS citations
RETURN d.name, d.id_icpsr, citations
ORDER BY citations DESC
```
````

Listing 6.5: Data and Metadata Search prompt

Here, to ensure consistency in the format of generated outputs and to accommodate the Cypher query output in the above prompt, we set the maximum try parameter for the LLM to three. This mechanism leverages the LLM's generative capabilities and provides potentially useful metadata query results for the user's reference. We also set a limit to output the top 10 relevant results, to ensure a balance between the readability of generated results and the coverage of truly useful metadata[5].

6.3.2 Discussion of the Ask Data Feature

The integration of directed discovery with IHSKG represents a significant advancement over traditional search methods, addressing several critical dimensions of data discovery in academic research. By leveraging the interconnected nature of SKGs, the GAUDS system aligns closely with the CEVI principles of Connectivity, Efficiency, Visibility, and Interactivity. Below, we detail how this integration advances each of these principles:

Connectivity in the context of GAUDS refers to the seamless integration of various data elements and their attributes within a single, coherent framework. Directly asking questions to the schema of IHSK enables a dynamic linking of datasets, publications, and metadata, creating a rich tapestry of interconnected information. This connectivity allows researchers to traverse from one node to another—be it a dataset linked to a particular author or a research output connected to specific funding details—thereby uncovering hidden relationships and fostering interdisciplinary research opportunities.

Efficiency in data discovery is significantly enhanced through the use of SKGs due to the reduction in time and effort required to locate and understand relevant datasets. Directed discovery via IHSKG utilizes intelligent querying mechanisms that anticipate and adapt to user needs by generating context-aware Cypher queries. These queries are tailored to extract the most relevant information based on the user's input, thereby reducing the cognitive load on researchers and minimizing the occurrence of irrelevant search results.

Visibility of datasets within the scholarly community is crucial for ensuring that valuable research is discovered and utilized. IHSKG improve the visibility of datasets by indexing them in a manner that emphasizes their relationships to other research outputs and metadata, which enables the cross entity searching, such as co-citation of datasets. This not only helps in surfacing lesser-known datasets that might be relevant due to their connection to widely recognized research but also enhances the discoverability of new and ancillary data points that could lead to innovative research hypotheses.

Interactivity in the GAUDS system is facilitated through an interface that responds dynamically to user queries. The use of LLMs to interpret and generate queries based on the schema of the IHSKG allows for a more engaging user experience. Researchers can interact with the system in a conversational manner, posing questions and refining their searches in real-time, which mirrors the intuitive nature of human inquiry and learning.

[5] We choose 10 here for convenience based on a simple interface length assessment. Since the related metadata, different from the Top K datasets, is not our research focus, we do not include it as a hyperparameter for tuning.

6.4 Find Data: Recommendation with Vector Search

The third option for research data discovery in the GAUDS system is through vector search in IHSKG. In this section, we discuss the methods for creating the vector search application and analyze the technical choices behind.

6.4.1 Methods

We first introduce the method for creating the optimal find data function through the vector search feature. We use as introduced in Chapter 5 and conducted hyperparameter tuning based on the recall metric.

A 3-dimension Hyperparameter Tuning

In our vector search hyperparameter tuning, we explore the impacts of different settings on the efficiency and effectiveness of retrieving dataset recommendations. The parameters tuned include vector index type, vector similarity threshold, and the number of top results (Top K) to return. We use a specific Cypher query as demonstrated below to interact with the vector search capabilities of our data recommendation system:

```
WITH {this_embedding} AS inputEmbedding
CALL db.index.vector.queryNodes('{embedding}', {top_k}, inputEmbedding)
YIELD node AS similarDATASET, score
WHERE score > {threshold}
RETURN similarDATASET
```

Listing 6.6: Hyperparameters for Vector Search in IHSKG

The parameter settings in our tuning include:

- Vector Index Type: Options are 'name', 'terms', and 'name and terms'. This choice determines the kind of textual information from the datasets that is embedded into vectors for similarity comparisons.
- Vector Similarity Threshold: We explore thresholds at 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. This parameter sets a cut-off for the similarity scores—only datasets with a similarity score above this threshold to the query vector are considered.
- Top K: Varying from 5 to 30 in increments of 5, this parameter specifies the number of top results to retrieve, focusing on those that are most similar to the input embedding.

Using both a "Top K" parameter and a "threshold" parameter serves to efficiently and effectively narrow down the search results based on similarity, ensuring both relevance and quality of the results. On the one hand, the "Top K" parameter specifies the number of top results to retrieve from the vector search based on the closest cosine similarity to the query vector. This parameter helps in limiting the search to a manageable number of results that are most similar to the input embedding. It provides a way to control performance and relevance, as retrieving and processing only the Top K results can significantly reduce the computational load and speed up the query, especially in large datasets. On the other hand, the "threshold" parameter sets a minimum similarity score that results must meet to be considered relevant. This score is a cut-off point for deciding which nodes are sufficiently similar to the input embedding. Using a threshold filters out less relevant results even within the Top K matches. This ensures that the quality of the results is maintained, as only those results that have a similarity score above the specified threshold are returned. It is particularly useful when the ”Top K” results include items of varying and possibly low similarity.

Evaluation

For the evaluation of our data recommendation system, we focus predominantly on recall, given the nature of the task, which is to identify suitable datasets for a given research query from a possibly vast repository. These metrics, Mean and Variance of Recall, allow us to assess not only how well the system performs on average but also how its performance might vary across different scenarios, ensuring we have a robust system capable of adapting to the query vector search needs.

Mean of Recall is calculated to evaluate the average effectiveness of the system across multiple queries in retrieving all relevant datasets. It is given by the following formula:

$$\text{Mean of Recall} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}\right)$$

where, N is the number of queries tested, TP_i are the relevant datasets correctly identified by the system for the i -th query (True Positives), and FN_i are the relevant datasets not retrieved by the system for the i -th query (False Negatives).

Variance of Recall helps us understand the consistency of the recall across different queries, indicating the stability of the model under various conditions. It is calculated as follows:

$$\text{Variance of Recall} = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} - \text{Mean of Recall}\right)^2\right)$$

6.4.2 Results

Figures 6.6 shows the exploration of hyperparameters within the vector search mechanism elucidates the intricate balance between achieving high recall and maintaining consistency in dataset retrieval. Comparing across all three types of indexes, when the threshold parameter decrease from 0.9 to 0.6, there is no further recall increase beyond it, indicating a threshold of 0.6, across all three indexes is the best.

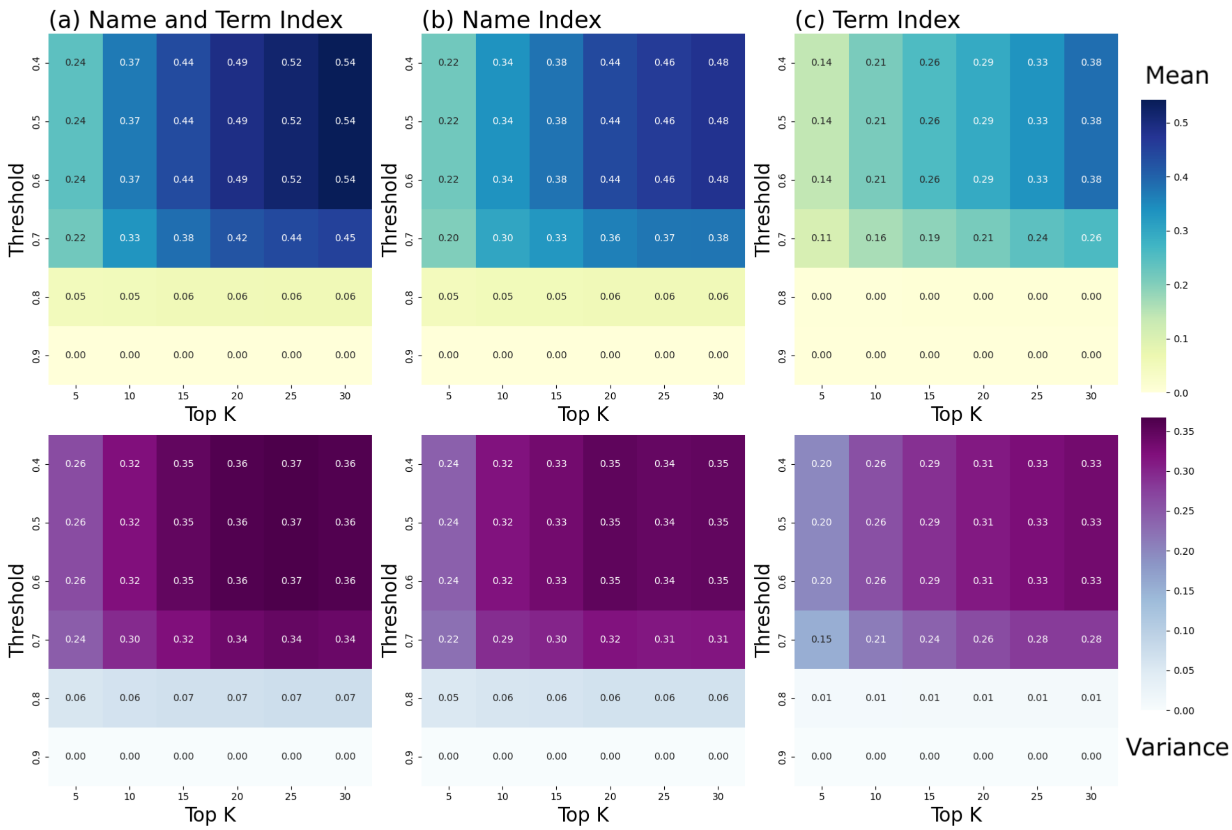

Figure 6.6: Data Citation Recall

For the Name and Term Index, the mean recall scores are indicative of a nuanced interplay between the Top K and threshold parameters. Specifically, at a threshold of 0.6, the mean recall scores stabilize around 0.24 to 0.54 as Top K values increase, suggesting a robust retrieval performance without significant trade-offs in precision. This range is maintained up to a Top K of 30, which supports the selection of K=30 as it suggests comprehensive coverage without compromising relevance.

For the Name Index and Terms Index at a threshold of 0.6, the recall increases when Top K gets larger. However, their highest recalls are 0.48 and 0.38 respectively, both of which are less than 0.54, showing that Name and Term Index remains the best index to use for the similarity search of publication embedding.

The variance of recall scores, which reflects the consistency of the search results, further informs the selection of hyperparameters. For the Name and Term Index, we observe a lower variance with a Top K of 30 across all thresholds, suggesting a stable recall performance. Similarly, the Name Index and Term Index both demonstrate manageable variances at a threshold of 0.6 with a Top K of 30, all around 0.36, implying that the results' reliability does not falter even as the breadth of the search increases.

6.4.3 Discussion of the Find Data Feature

Choosing a threshold of 0.6 strikes a harmonious balance between inclusivity and precision. It is stringent enough to ensure relevance in the search results but not so restrictive as to omit valuable datasets. Meanwhile, a Top K of 30 is indicative of an expansive search scope, capturing a comprehensive array of datasets without overwhelming the researcher with an excessive number of options. Together, these parameter choices aim to provide a thorough, consistent, and relevant selection of datasets for researchers delving into comprehensive datasets repositories. In addition to the threshold, we would also like to discuss the nuances in the choices of vector indexes and the Top K values, where we combine the experimental results' implications with user-centric requirements.

Vector Index: Datasets' Name versus Terms. The choice of vector index—a critical component in the vector search mechanism—revolves around whether to utilize datasets' names, terms, or a combination of both. Dataset names often contain concise descriptors that can quickly align with user queries, while terms within the datasets may provide a richer context that encapsulates the intricacies of the dataset content. The observed mean recall scores indicate that a composite approach, integrating both names and terms, leverages the brevity of names and the comprehensive context provided by terms. This dual-indexing strategy results in a higher probability of matching a user's search intent with relevant datasets, as evidenced by the enhanced mean recall outcomes. Conversely, indexing solely on names or terms may result in a narrower scope of search and potentially overlook pertinent datasets that could be discovered through the more inclusive, hybrid indexing approach.

Top K: Practical versus Theoretical Upper Bounds. The determination of the Top K parameter, representing the number of top results retrieved, oscillates between practical utility and theoretical exhaustiveness. While a higher Top K theoretically offers a more exhaustive review of potential datasets, it risks inundating users with less relevant results, thereby diluting the quality of recommendations. Empirical data from the heatmaps suggest that beyond a certain threshold, increments in Top K do not significantly enhance mean recall, indicating a point of diminishing returns. A Top K of 30 emerges as a pragmatic upper bound that encompasses a breadth of options without encroaching on the territory of impracticality. More practically, it is not possible for users to review a long list of publications, while the number of 30 is likely the upper bound for a small number of cursor scrollings to finish the result reviewing. This boundary reflects a convergence of computational efficiency and user manageability, ensuring the retrieved datasets remain within the realms of relevance and practical examination for the researcher.

6.5 Reuse Guide Generation and Visualization

In this section, we explain the process of ensuring that datasets are not only accessible but also aligned with the specific needs of research queries. After the user finds the datasets they are interested in for further exploration and generation of reuse guides, they input the dataset IDs to obtain further guidance on how to use them[6]. By employing a heuristic approach grounded in practical examples, we aim to illustrate the essential

[6]The ID input feature can potentially be replaced in later versions by a selection feature in the tabulated results. We discuss more about the known limitations of the design in the alpha version of GAUDS in Appendix B.

considerations and methodologies that guide researchers in selecting and utilizing datasets effectively. For instance, consider a researcher querying:

```
I'd like to do research related to cardiovascular disease. What datasets should I use?
    How should I use them?
```

Listing 6.7: Resue Guide Example Natural Language Query

This natural language query serves as a practical example for our discussion, guiding us through three critical subsections: **Fit for Purpose**, where we evaluate how well a dataset meets the specified research needs; **Data Documentation**, which emphasizes the importance of detailed dataset descriptions for effective use; and **Reuse Guide Visualization**, where we explore visual tools that enhance the understanding of dataset connections and contexts. Each of these subsections builds upon the initial query to provide a comprehensive framework for dataset selection and application, ensuring researchers can leverage data to its full potential.

6.5.1 Fit for Purpose

Ensuring that data is fit for the intended research purpose necessitates a meticulous selection process and it often through meticulous manual process by data managers [311]. The concept of ”fit(ness) for use” underscores the importance of data reliability and relevance in meeting specific research needs [312], [313]. The qualitative and experience-based measurement of reliability and relevance is widely utilized to assess dataset fit for purpose in both social science and health and medical research domains, for example, including public intrusion [314], government health information [315], and marine biodiversity [316].

In the GAUDS system, similarly, we focus on the assessing the reliability and relevance of the dataset to the user query, where we operationalize the core criteria in Structured Process to Identify Fit-For-Purpose Data (SPIFD) [313] as ”usefulness, accuracy, completeness, and reliability.” We use these concepts in the prompt to generate fit for purpose assessment and leverage LLMs' generative nature to contextualize the answers with respect to the user query and output granular potential research questions. The generated outputs include aspect-based reliability and relevance for both positive and negative potential research directions, which guide the data users to understand the full potential of the datasets and make informed data reuse decisions. We provide the detailed prompts that pass in the dataset name, the user question, and the respective generation heuristics as follows.

```
# Positive Aspect
Based on the dataset name and the user question below, what are the potential research
    questions that can be answered using this dataset? The criteria for assessing
    positive aspects of data fitness encompass finding usefulness of the dataset and
    suggestion detailed research questions, contextualized to the user question at hand
    .
\n Dataset name: {}
\n User question: {}
\n Output the suggested research questions and reasoning by point.

# Negative Aspect
```

```
Based on the dataset name and the user question below, what are the limitations of the dataset in answering the user's questions? The criteria for assessing negative aspects of data fitness encompass accuracy, completeness, and reliability, based on the possible research directions derived from the user's natural language query.
\n Dataset name: {}
\n User question: {}
\n Output the suggested research questions and reasoning by point.
```

Listing 6.8: Fit for Purpose Prompt

As such, we provide information and LLM-assisted evaluation to assess if the data aligns with the user's natural language query for research data needs. The criteria for assessing data fitness encompass related questions for positive aspects of fitness, as well as accuracy, completeness, and reliability concerns for negative aspects of fitness. The example below shows the Fit for Purpose guide generated for the query about cardiovascular disease and the dataset "Clinical Database to Support Comparative Effectiveness Studies of Complex Patients, 2005-2010 [United States]".

```
# Positive Aspect
Potential research questions that can be answered using the dataset "Clinical Database to Support Comparative Effectiveness Studies of Complex Patients, 2005-2010 [United States]" related to cardiovascular disease:

1. What are the comparative effectiveness of different treatment options for cardiovascular disease in complex patients?
- This dataset likely contains information on various treatment options and outcomes for patients with cardiovascular disease, allowing for a comparison of their effectiveness.

2. How do comorbidities impact the outcomes of cardiovascular disease treatment in complex patients?
- The dataset may include information on comorbidities and their effects on the treatment and outcomes of cardiovascular disease in complex patients.

3. What are the long-term outcomes of different interventions for cardiovascular disease in complex patients?
- By analyzing the data over a period of 5 years (2005-2010), researchers can assess the long-term outcomes of different interventions for cardiovascular disease in complex patients.

4. How do demographic factors such as age, gender, and race influence the management and outcomes of cardiovascular disease in complex patients?
- The dataset may contain demographic information that can be used to analyze how factors such as age, gender, and race impact the management and outcomes of cardiovascular disease in complex patients.

5. What are the trends in the prevalence and management of cardiovascular disease in complex patients over the study period?
```

```
- By examining the data from 2005 to 2010, researchers can identify trends in the
    prevalence and management of cardiovascular disease in complex patients, providing
    valuable insights for future research and healthcare practices.

# Negative Aspect
Suggested research questions:
... (repeat the research questions above)

Limitations of the dataset in answering the user's questions:
1. The dataset covers a specific time period from 2005-2010, which may not provide the
    most up-to-date information on cardiovascular disease treatments and outcomes.
2. The dataset focuses on complex patients, which may not represent the general
    population with cardiovascular disease.
3. The dataset may not have specific variables related to cardiovascular disease
    treatments and outcomes, limiting the ability to conduct detailed analyses on this
    topic.
4. The dataset may not include information on specific cardiovascular disease subtypes
    or risk factors, which could impact the relevance of the findings to the user's
    research interests.
```

Listing 6.9: Fit for Purpose Result Example

For the positive aspects, the suggested research questions generated from the prompt demonstrate the dataset's potential utility in exploring various aspects of cardiovascular disease treatment and outcomes in complex patients. These questions focus on the comparative effectiveness of treatment options, the impact of comorbidities, long-term outcomes, demographic influences, and trends over the study period. The reasoning provided highlights the dataset's expected content, such as treatment details, outcomes, comorbidities, and demographic data, which are essential for such analyses. This kind of detailed breakdown ensures that researchers can plan robust studies with clear expectations of what the dataset can deliver, enhancing the research's relevance and precision.

For the negative aspects, conversely, the outputs provide critical insights into the dataset's limitations, helping to set realistic expectations for potential users. The limitations mentioned include the dataset's confined time scope (2005-2010), which may not reflect the latest advancements in cardiovascular disease treatment. Additionally, the focus on complex patients might limit generalizability to the broader population of cardiovascular disease sufferers. Another significant limitation is the possible absence of specific variables or detailed information on subtypes and risk factors of cardiovascular disease, which could affect the comprehensiveness of the research findings.

Overall, the the results generated by the Fit for Purpose prompt can effectively guide researchers in evaluating the appropriateness of a dataset for specific research queries. This guidance is vital for ensuring that the selected data aligns well with the research objectives, thereby maximizing the efficacy and impact of the research conducted. This process aims to support the integrity and reliability of research findings. This dual analysis approach helps researchers in several ways:

- Informed Decision-Making: Users, especially researchers, can better decide whether this dataset is appropriate for their specific research questions, based on a clear understanding of what the data can and cannot provide.

- Research Planning: By knowing the dataset's strengths and weaknesses, researchers can tailor their methodologies and analysis techniques to accommodate or compensate for these factors, potentially by combining this dataset with other data sources.
- Expectation Management: Understanding both the positive potential and the limitations helps manage expectations regarding the research outcomes and the conclusions that can be robustly supported by the data.
- Resource Optimization: Researchers can optimize their use of time and resources by focusing on research questions that the dataset is well-suited to address, avoiding areas where the dataset falls short.

6.5.2 Data Documentation

Data documentation and metadata play pivotal roles in the comprehension, interpretation, and subsequent reuse of research data. Comprehensive data documentation provides critical context, detailing the data collection methodology, structure, manipulations, and any conditions related to data confidentiality and access. Metadata, or data about data, enriches this context by detailing the origin, purpose, and conditions of use, facilitating data discovery, indexing, and citation. Adherence to established metadata standards enhances data interoperability and findability, crucial for fostering data sharing and reuse within and across research communities. The following prompt show the six aspects, including data owner, data funder, related top publications, related top data users, related datasets, and if the dataset is restrcted.

```
# write query for finding the owner: dataset <-HOST- owner
query_owner = f"MATCH (d:DATASET)<-[:HOST]-(o) WHERE d.id_icpsr = {this_dataset} RETURN
     o as owners"

# write query for finding the funder: dataset <-SUPPORT- funder
query_funder = f"MATCH (d:DATASET)<-[:SUPPORT]-(f) WHERE d.id_icpsr = {this_dataset}
    RETURN f as funders"

# write query for top 5 most cited publications that use this dataset
query_most_cited = f"MATCH (d:DATASET)<-[:CITE]-(p:PUBLICATION) WHERE d.id_icpsr = {
    this_dataset} RETURN p, COUNT(p) AS citations ORDER BY citations DESC LIMIT 5"

# write query for top 5 most frequent data user authors (through publications)
query_most_author = f"MATCH (d:DATASET)<-[:CITE]-(p:PUBLICATION)<-[:WRITE]-(a:AUTHOR)
    WHERE d.id_icpsr = {this_dataset} RETURN a, COUNT(a) AS publications ORDER BY
    publications DESC LIMIT 5"

# write query for the 5 related datasets (through publications) through the concept of
    "co-citation"
```

```
query_related = f"MATCH (d:DATASET)<-[:CITE]-(p:PUBLICATION)-[:CITE]->(d2:DATASET)
    WHERE d.id_icpsr = {this_dataset} RETURN d2.id_icpsr, d2.name, COUNT(p) AS
    citations ORDER BY citations DESC LIMIT 5"
```

Listing 6.10: Data Documentation Prompt

With the above six dimensions of data documentation, we provide users with the metadata related to the data collection, access, previous reuse, which prepare the user to start the research on the data conditions and the corresponding data manipulation methods. Users can also evaluate if the provided data follows community standards, and researchers from which community have used it, for data formatting and sharing, potentially helping streamline data integration and reuse by ensuring interoperability and reducing data cleaning efforts.

For instance, here are the results generated with the query about cardiovascular disease and the dataset ”National Health Interview Survey, 1994: Year 2000 Objectives Supplement”[7]. The generated data documentation uses a list of metadata information that extended the currently available information on the ICPSR Find Data webpage. The structured queries and subsequent data documentation provide a clear, multifaceted view of the dataset's usage, impact, and connectivity within the research community[8]. Further improvements in frontend web design could enhance the user experience. Nonetheless, the existing comprehensive data documentation effectively supports data discoverability and reuse, emphasizing the significance of complete and high-quality metadata in facilitating efficient and effective research practices

```
"Owner": [
{
  "owners": {
    "name": "NACDA"
  }
}
],

"Funder": [],

"Most cited publications": [
{
  "p": {
    "year": 2000,
    "name": "Mobility impairments and use of screening and preventive services",
    "id_openalex": "https://openalex.org/W1604876104",
    "related_works": [list omitted for simplicity],
    "datasets": 6875,
    "reference_count": 339
  },
  "citations": 1
},
...
],
```

[7]We provide the full results in Appendix B.3

[8]In the alpha version, notably, we directly demonstrate the JSON file to users with folding options to hide the viewed sections, while we plan to use tabular formats in later versions to provide a clearer view to the us Appendix B.

```

"Most frequent author": [
{
  "a": {
    "author_id": "https://openalex.org/A5051038042"
  },
  "publications": 3
},
...
],

"Related datasets": [
{
  "d2.id_icpsr": 6344,
  "d2.name": "National Health Interview Survey, 1992: Cancer Control Supplement",
  "citations": 5
},
...
],

"Restriction": [
{
  "d.restriction": "AVAILABLE"
}
]
```

Listing 6.11: Data Documentation Example Output (Selected)

Data Ownership and Funding From the generated documentation, we identify the National Archive of Computerized Data on Aging (NACDA) as the data owner. This highlights the role of NACDA in hosting and managing the dataset, which is crucial for ensuring the dataset's availability and integrity over time. The absence of a listed funder in the results indicates that specific funding sources either were not disclosed or were not directly linked to the dataset through the recorded metadata. This could suggest a need for enhanced funding disclosure practices to better acknowledge financial support and its impact on research resources.

Publications and Usage Patterns The results also include details about the most cited publications that utilized the dataset. Each listed publication has only one citation, which may indicate either a niche area of study or underutilization of the dataset within the broader research community. Notably, the topics covered by these publications span diverse health issues, from mobility impairments to occupational health, demonstrating the dataset's versatility and applicability across multiple public health inquiries.

Data User Engagement The analysis of the most frequent data user authors reveals that several researchers have multiple publications citing this dataset, but the overall frequency is relatively low (maximum three publications per author). This suggests a concentrated use among a small number of researchers rather than widespread engagement across the community. Recognizing these authors could be advantageous for new researchers looking to collaborate or seek expert insights.

Related Datasets and Interoperability The identification of related datasets through co-citation analysis is particularly useful for researchers aiming to conduct comparative studies or meta-analyses. The related datasets, such as the NHIS supplements from various years (shown in Appendix B), suggest a continuity in the research topics and potential for longitudinal studies. The similarity and repeated citation of related datasets underscore their relevance and utility in ongoing public health research, facilitating data interoperability and thematic analysis over time.

Data Restrictions and Access Finally, the dataset is marked as ”AVAILABLE,” indicating no significant restrictions on its access. This accessibility is critical for fostering an open research environment where datasets can be freely used and reused, thus enhancing scientific discovery and transparency[9].

6.5.3 Reuse Guide Visualization

After the user input the dataset IDs and obtain the reuse guide, they can then obtain the interactive visual guidance (by clicking on the Show Table & Graph button), including a table with links to the datasets they choose and the metadata nodes and edges related to those datasets. We will provide an example in the case study section next and more examples in Appendix B.

6.6 Case Studies

In this section, we first continue with the query example about cardiovascular disease in the previous section, focusing on introducing the overall workflow of the system. We then present an additional case on dementia research, highly the linked resources from the system. These two cases together showcase the utilities and impacts of the GAUDS system.

[9] ICPSR provide further explanations for their availability notes and shows the way they manage and grant access to restricted data: https://www.icpsr.umich.edu/web/pages/ICPSR/access/restricted/.

6.6.1 Case 1: Cardiovascular Disease Research

Explore (by topics). In the Explore feature, by entering a general interest topic, such as cardiovascular diseases, the user can receive a curated list of relevant datasets from the IHSKG. Figure 6.7 displays the interface after a user selects 'Explore' and inputs 'Cardiovascular Disease' into the query box, showcasing datasets categorized under relevant topics such as 'Cardiology and Cardiovascular Medicine' and 'Public Health'.

Topics:


```
[
  0 : "Cardiology and Cardiovascular Medicine"
  1 : "Public Health, Environmental and Occupational Health"
]
```


Top datasets by topic:


```
[
  0 : {
    "d.name" : "Charleston Heart Study, Charleston, South Carolina, 1960-2000"
    "d.id_icpsr" : 4050
  }
  1 : {
    "d.name" :
    "National Health Examination Survey, Cycle I, 1959-1962: Cardiovascular
    Findings"
    "d.id_icpsr" : 9206
  }
  2 : {
    "d.name" :
    "Evaluation of the Psychological Effects of Administrative Segregation in
    Colorado, 2007-2010"
    "d.id_icpsr" : 31321
  }
  3 : {
    "d.name" : "Galveston Bay Recovery Study, 2008-2010"
    "d.id_icpsr" : 34801
  }
```


Figure 6.7: Example Search Results by Clicking 'Explore'

Ask (database metadata). The Ask functionality allows users to query database metadata directly. As illustrated in Figure 6.8, when a user inputs specific questions regarding data availability or metadata characteristics, the system provides detailed responses that include dataset names, terms associated with the datasets, restrictions, and other pertinent metadata. This facilitates a deeper understanding and quick access to specific dataset information.

Results based on your query question:


```
[
  0 : {
    "d.name" : "Charleston Heart Study, Charleston, South Carolina, 1960-2000"
    "d.id_icpsr" : 4050
    "d.terms" :
    "activities of daily living; African Americans; aging; cardiovascular
    disease; death records; depression (psychology); health status; mental
    health; physical condition; psychological wellbeing; race; social behavior;
    White Americans"
    "d.restriction" : "AVAILABLE"
    "d.year" : "neo4j.time.DateTime(2005, 1, 28, 5, 0, 0, 0, tzinfo=<UTC>)"
    "d.date" : "neo4j.time.DateTime(2005, 1, 28, 5, 0, 0, 0, tzinfo=<UTC>)"
  }
  1 : {
    "d.name" :
    "National Health Examination Survey, Cycle I, 1959-1962: Cardiovascular
    Findings"
    "d.id_icpsr" : 9206
    "d.terms" :
    "cardiovascular disease; chronic illnesses; dental health; eyesight; health
    behavior; health history; health status; hearing (physiology); medical
    evaluation; physical health; testing and measurement"
    "d.restriction" : "AVAILABLE"
    "d.year" : "neo4j.time.DateTime(1989, 9, 26, 4, 0, 0, 0, tzinfo=<UTC>)"
    "d.date" : "neo4j.time.DateTime(1989, 9, 26, 4, 0, 0, 0, tzinfo=<UTC>)"
  }
  2 : {
    "d.name" :
```


Figure 6.8: Example Search Results by Clicking 'Ask'

Find (via vector search). The Find feature employs vector search technology to match the query's semantic context with datasets that are most topically relevant. Figure 6.9 shows the results for a vector search based on the query "Cardiovascular Disease." It lists datasets that are semantically closest to the query, thus ensuring that the results are highly targeted and contextually appropriate for the user's research needs.

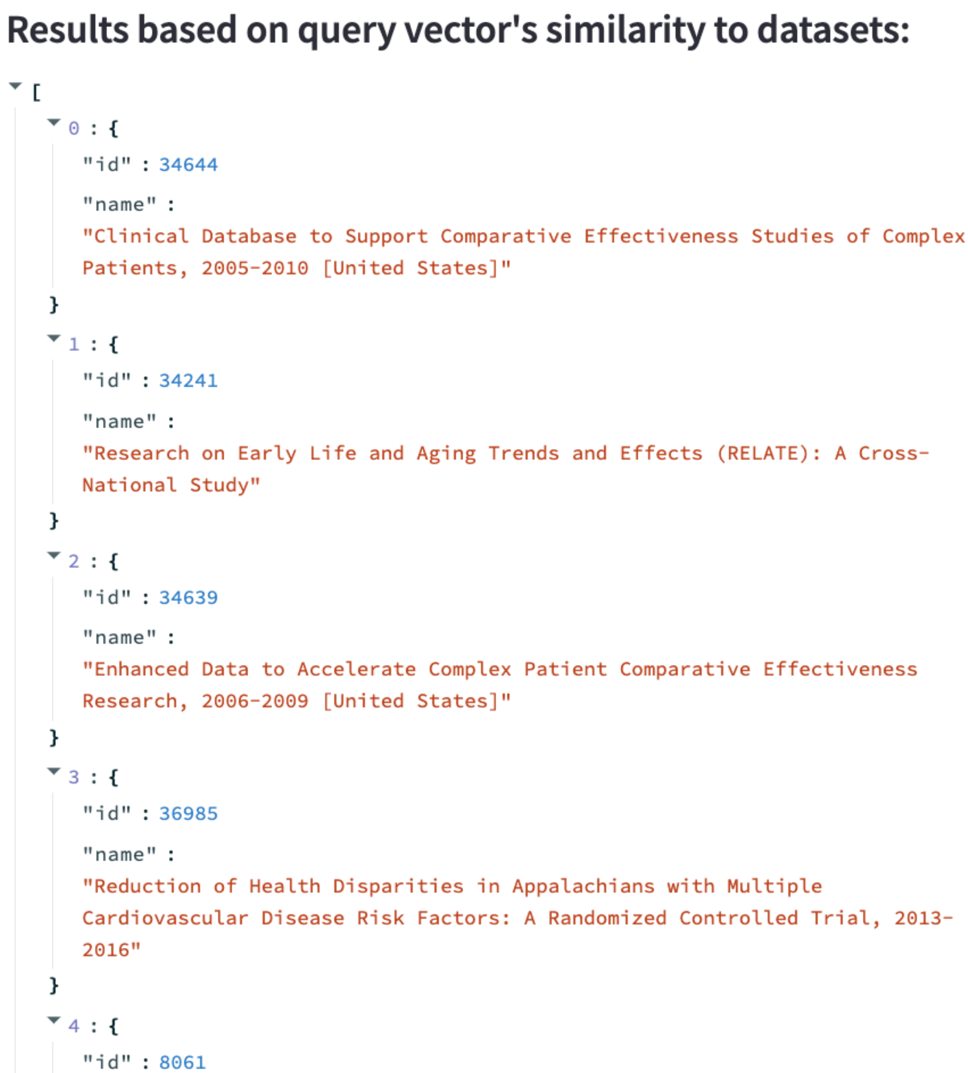

Figure 6.9: Example Search Results by Clicking 'Find'

Reuse Guide Generation and Visualization. After identifying relevant datasets, users can generate a Reuse Guide for each dataset to facilitate effective data application and integration. This guide provides essential details such as 'Fit for Purpose' and 'Data Documentation.' Additionally, the system allows for the visualization of relationships (with category marked on edges) between datasets, authors, and publications through interactive graphs and tables. Figure 6.10 exemplifies the GAUDS system frontend displaying a Reuse Guide along with the interactive network graph and table, illustrating the comprehensive data connections.

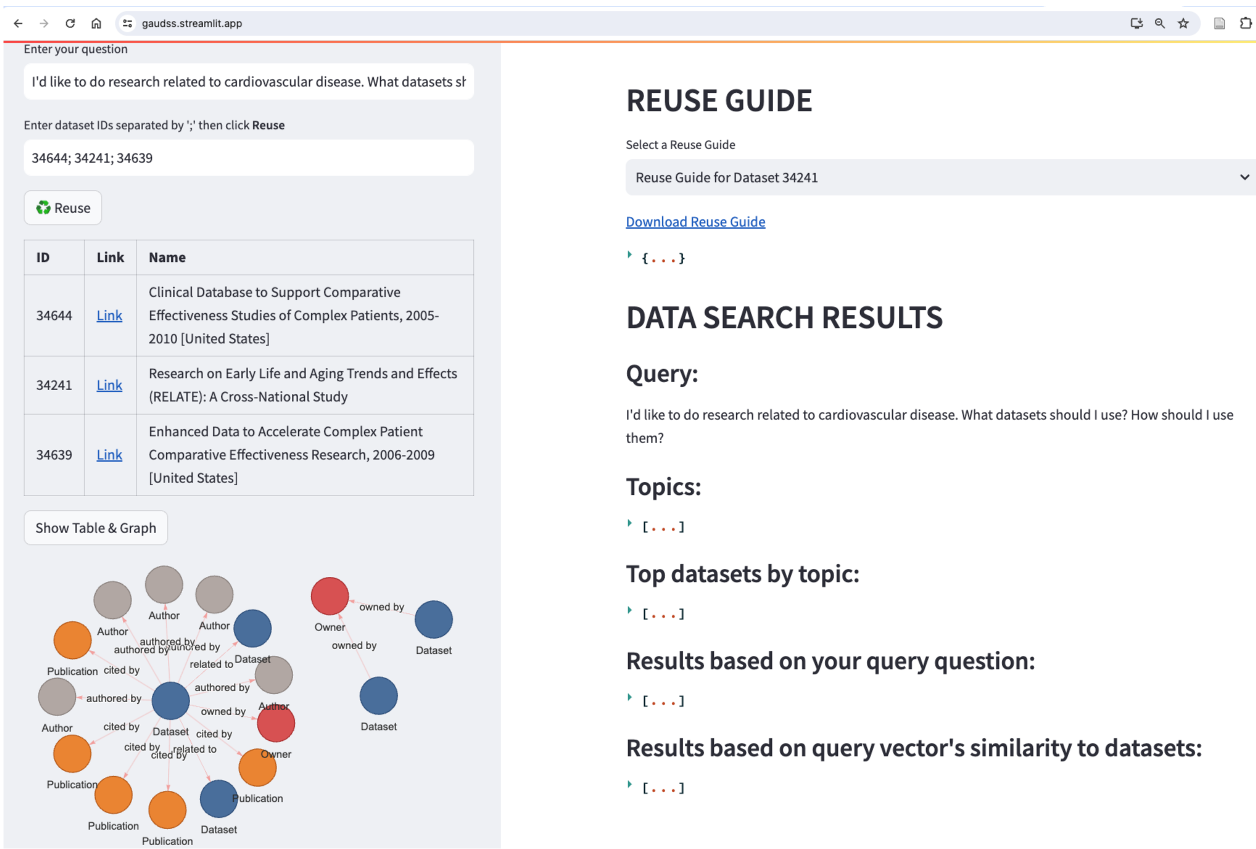

Figure 6.10: GAUDS System Frontend with an Example of Search Results

6.6.2 Case 2: Dementia Research

For the dementia research example, we use the first query from the second focus group study participant, which will be introduced in detail in Chapter 7. The natural language query is as follows.

```
What datasets are available for studying dementia in populations over 65 years old?
```

Listing 6.12: PQ 2.1

The Table & Graph of PQ 2.1

Figure 6.11 shows the search results and the selected linked web pages. Participant 2 chooses three datasets to study dementia, including 2877, 3417, and 36589 (the names and details of the datasets are presented in the table in Figure 6.11 (a)). There are also related entities visualized in the side bar, including 1) the data owners, the authors who used the datasets, and the publications that cited the datasets. The related datasets are also presented, including ”Aging of Veterans of the Union Army: United States Federal Census Records, 1850, 1860, 1900, 1910” (Dataset 6836), ”Aging of Veterans of the Union Army: Military, Pension, and Medical Records, 1820-1940” (Dataset 6837), ”United States Census of Mortality: 1850, 1860, and 1870” (Dataset 2526), and ”National Health and Nutrition Examination Survey III, 1988-1994”.

We also provide examples of a data reuse author and a data citing publication, linked to their OpenAlex pages. Figure 6.11(b) shows that the author Chulhee Lee has published in various health and medical domains and their data are used work can provide more examples to the user. Figure 6.11(c) shows that the publication by Dora L. Costa used the dataset and the user can further investigate the publication metrics by clicking on the OpenAlex web page.

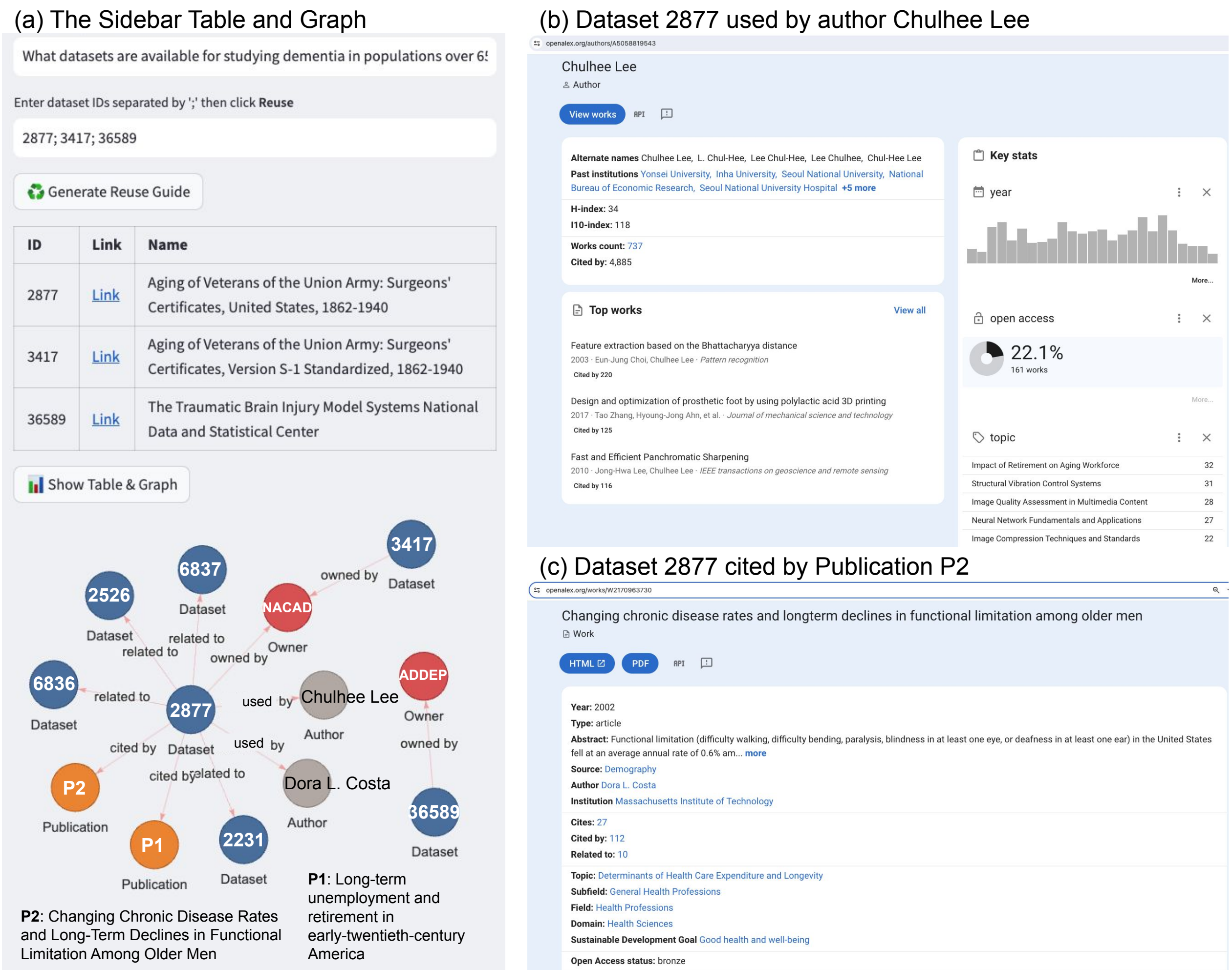

Figure 6.11: Reuse Guide Table and Graph Details of PQ 2.1

The Fit for Purpose Details of PQ 2.1 and Dataset 2877

The following Fit for Purpose analysis show the generated results for the query ”What datasets are available for studying dementia in populations over 65 years old?” regarding dataset 2877. It contextualizes the data reuse suggestion to the research query and suggests secondary research questions, as well as how to use the dataset in each of the questions. The fit for purpose analysis is generated according to both the natural language query on the topic of dementia for population over 65 years of age and the specific dataset regarding veterans. The generated results successfully understand the combined topics from the input and the database.

```
Fit for Purpose (Positive Aspects):
Potential research questions that can be answered using the dataset "Aging of Veterans of the Union Army: Surgeons' Certificates, United States, 1862-1940" in relation to studying dementia in populations over 65 years old could include:

1. What were the prevalence rates of dementia among veterans of the Union Army as they aged from 65 years old and older? This dataset provides information on the health and aging of veterans over a long period of time, making it possible to track the prevalence of dementia in this population.

2. How did factors such as wartime experiences, injuries, or medical treatments received during their service impact the development of dementia in veterans of the Union Army? By analyzing the data on surgeons' certificates, researchers can explore potential risk factors for dementia in this specific population.

3. What were the common symptoms and progression of dementia among veterans of the Union Army as documented by surgeons over time? The dataset likely contains detailed medical records that can provide insights into the symptoms and progression of dementia in this population.

4. How did socioeconomic factors, such as income level or access to healthcare, influence the prevalence and outcomes of dementia among veterans of the Union Army? By examining the demographic and socioeconomic information available in the dataset, researchers can investigate the impact of these factors on dementia in older populations.

Overall, this dataset offers a unique opportunity to study the aging process and health outcomes of a specific population over time, providing valuable insights into the prevalence, risk factors, and outcomes of dementia in older individuals.

Fit for Purpose (Negative Aspects):
Limitations of the dataset in answering the user's question:

1. Time Period: The dataset covers the years 1862-1940, which may not be relevant to studying dementia in populations over 65 years old in more recent times.

2. Specific Focus: The dataset specifically focuses on the aging of veterans of the Union Army and surgeons' certificates, which may not provide comprehensive information on dementia in the general population over 65 years old.
```

```
3. Lack of Specific Variables: The dataset may not have specific variables related to dementia or cognitive decline, making it difficult to directly study this condition in the population.

4. Limited Sample Size: The dataset may have a limited sample size of veterans of the Union Army, which may not be representative of the general population over 65 years old.\n

5. Missing Data: There may be missing or incomplete data in the dataset, which could affect the accuracy and reliability of any findings related to dementia in populations over 65 years old.

Suggested research questions and reasoning:

1. What are the risk factors for dementia in populations over 65 years old in the present day? This question would require a dataset with more recent data and specific variables related to dementia risk factors.

2. How does the prevalence of dementia vary among different demographic groups over 65 years old? This question would require a dataset with a larger and more diverse sample size to provide more representative results.

3. What are the trends in the diagnosis and treatment of dementia in populations over 65 years old over time? This question would require a dataset with longitudinal data on dementia diagnosis and treatment, which may not be available in the current dataset.

4. How does the presence of comorbidities impact the progression of dementia in populations over 65 years old? This question would require a dataset with detailed medical history and health information, which may not be present in the dataset on the aging of veterans of the Union Army.
```
Listing 6.13: Fit for Purpose Example for PQ 2.1 and Dataset 2877

6.6.3 Discussions of Utilities and Impacts

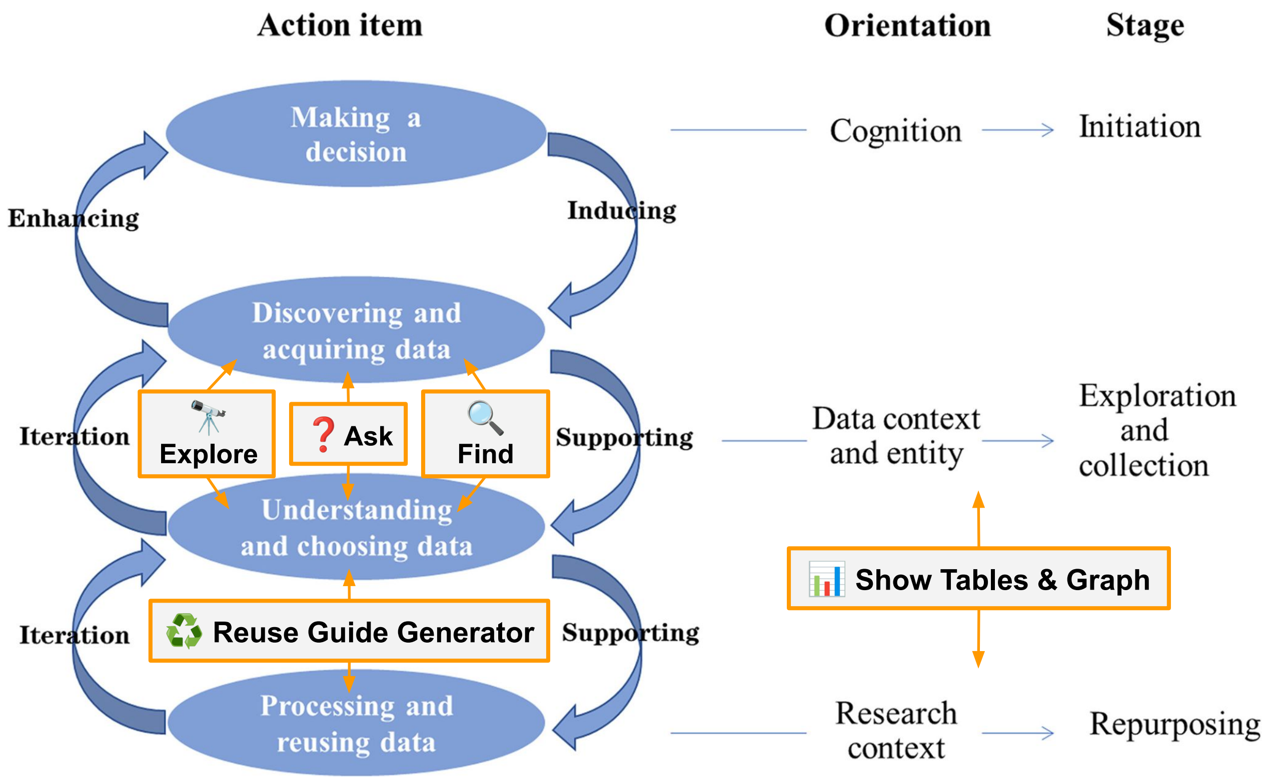

Figure 6.12: GAUDS in Data Reuse Process (adapted from [317])

The case studies of the GAUDS system, centered around the exploration of datasets related to cardiovascular disease, exemplify the system's robust capabilities in handling complex queries and facilitating comprehensive data discovery and utilization. The system's integration with IHSKG and advanced search functionalities such as 'Explore', 'Ask', 'Find', and 'Reuse Guide Generation' enable researchers to navigate through extensive data repositories with precision and efficiency. The system acts as a robust semantic agent, adept at managing the diversity and depth of academic datasets. The system can handle both semantic flexibility in research topics' representations and the complexity of database schema, highlighting the system's capabilities and the potential user satisfaction. Through careful choices and the well-documented adaptation of proprietary LLMs, the GAUDS system facilitates the transparent use of generative AI, which can potentially promote open supervision and responsible use of AI. In general, as Figure 6.12 indicates, the GAUDS system effectively implements CEVI principles. It connects research entities within the SKG, enabling effective searches characterized by flexibility and relevance. In addition, it displays search results through an accessible interface, enhancing visibility, and supports interactive interpretation of these results, allowing extensive customizability.

System Efficiency and Relevance. The 'Explore' functionality allows users to initiate their search by entering broad topics, which the system then contextualizes to present the most relevant datasets. This not only streamlines the discovery process but also ensures that the datasets presented are aligned with current research interests and topics. The ’Ask’ functionality further enhances this by providing direct access to detailed metadata, allowing researchers to quickly ascertain the suitability of datasets for their specific needs.

Semantic Precision and User Interaction. The ’Find’ feature, utilizing vector search technology, demonstrates the system’s ability to interpret and respond to the semantic nuances of user queries. This capability ensures that the search results are not only relevant but also contextually tailored to the user’s research intent, thus enhancing the overall search experience and outcome. The user interaction is intuitive and responsive, accommodating both novice and experienced researchers.

Implications for Information Retrieval (IR). The system’s design incorporates significant contributions to IR theory, particularly in balancing precision and recall. Precision refers to the accuracy of the search results in meeting the user’s query, while recall pertains to the system’s ability to retrieve all relevant items. The trade-offs between these metrics have implications for serendipity and discovery. High precision may limit the breadth of results, potentially missing out on unexpected yet valuable datasets. Conversely, a focus on recall may present users with a broader range of data, enhancing opportunities for serendipitous discoveries but potentially overwhelming them with less relevant results. The precision/recall trade-off is crucial in determining the effectiveness of the ’Explore’ and ’Find’ functionalities. For instance, a high precision approach ensures that the datasets shown are highly relevant to the user’s query, minimizing time spent sifting through irrelevant data. However, this may reduce the chance of uncovering unexpected, potentially groundbreaking datasets. On the other hand, a system with high recall may enhance serendipity by exposing users to a wider array of data, but at the cost of presenting more irrelevant information. Therefore, the system must strike a balance, optimizing for the user’s needs while maintaining an avenue for serendipitous discovery. This balance is particularly important in academic research, where both directed searches and unexpected insights can be valuable.

Data Reusability and Network Visualization. The ’Reuse Guide Generation’ is particularly noteworthy as it not only facilitates effective data utilization but also promotes data reusability among the research community. The guides provide detailed documentation and context, which are crucial for ensuring that datasets are used appropriately and to their fullest potential. Furthermore, the visualization of data relationships via interactive graphs and tables not only aids in understanding the data landscape but also in identifying potential new avenues for research.

Towards Transparency and Responsible Use of AI. The adaptation of general-purpose LLMs, originally trained on diverse and broad datasets from the web and social media, to specialized scholarly contexts such as the GAUDS system presents unique challenges and responsibilities. Table 6.3 shows the current and planned (marked in parentheses) features in the GAUDS system that aim to facilitate transparent and responsible use of AI for research data discovery and reuse.

Table 6.3: Towards Transparent and Responsible Use of LLMs in GAUDS

Feature	**Description**
Clear documentation of data sources, training methods and algorithms	Documentation to ensure transparency and user trust.
Model Fine-Tuning	Technical measure to adapt LLMs to specific academic fields, improve relevance and accuracy, and reduce hallucinations.
Mandatory user agreements (planned)	Ethical and Responsible Use Guidelines to ensure ethical use, protecting anonymity of personal and biological information.
Continuous Monitoring and Updating (planned)	Regular updates and monitoring to adapt to new research findings and maintain model performance.
Engaging with the academic community (planned)	Stakeholder Engagement to refine system functionalities and address emerging issues.

To ensure transparency in this system, we designed it with clear documentation of the data sources, training methods, and algorithms used, allowing researchers to understand and trust the system output[10]. The system also has the great potential to ensure the responsible use of AI technologies in scholarly settings through implementing specific technical and ethical guardrails. On the technical end, we leverage *Model Fine-Tuning* on scholarly datasets, ensuring that the model's responses are not only accurate but also relevant to specific academic fields. This process helps mitigate the issue of generating nonfactual content ("hallucinations") by contextualizing the model reasoning to academic and data search domains to maintain the integrity of research outputs. In terms of ethics, we plan to enforce a set of system use protocols for public access. When the system and the web search application is public online, we will have all users sign an agreement on *Ethical and Responsible Use Guidelines*. Due to its openness and well-formed documentation, the GAUDS system can incorporate a set of ethical guidelines to govern the use of AI technologies, depending on specific organizational needs. For example, the anonymity of subjects' personal identifiable information (PII) and biological information [318]–[320]. These guidelines, as with any newly developed guidelines, are aimed at preventing misuse and ensuring that AI is used in a manner that respects academic integrity and promotes responsible research practices. We will also conducting *Continuous Monitoring and Updating* and prompting *Stakeholder Engagement* for the system. Regular monitoring and updating of AI models will be conducted to adapt to new research findings and evolving academic standards. This ongoing maintenance helps prevent the degradation of model performance over time and ensures continued relevance and accuracy. Engaging with a diverse group of stakeholders from the academic community helps refine the functionalities of the system and ensure that it meets the nuanced needs of researchers. This feedback loop is essential for continuous improvement and for addressing emerging issues promptly.

[10]We will make the GAUDS system open source when submitting the corresponding journal paper.

Potential Impact on Research Efficiency. Overall, the GAUDS system significantly enhances research efficiency by reducing the time and required domain expertise for data discovery and preparation. Its emphasis on interactivity, coupled with the ability to visualize complex relationships within the data, fosters a user-friendly research data discovery and reuse environment. In particular, the GAUDS system profoundly enhances the data discovery and reuse process, addressing critical pain points across various stages of data discovery and reuse, including *Initiation*, *Exploration and Collection*, and *Repurposing* [317], [321], [322]. Figure 6.3 shows how features in the GUADS system enhance steps in the theoretical model of the data reuse process synthesized by [317].

The steps implemented in the GAUDS system parallel the process of a reference interview in a library setting. A reference interview involves a librarian helping a user articulate their information needs and guiding them to relevant resources. Similarly, the 'Explore' functionality of the GAUDS system acts like the initial stage of a reference interview, where broad topics are identified and contextualized to align with the user's research interests. As the interaction progresses, the 'Ask' functionality mirrors the clarifying questions in a reference interview, providing detailed metadata to refine the search and assess the relevance of datasets. Finally, the 'Find' feature corresponds to the latter part of a reference interview, where the librarian helps the user interpret and use the information. By utilizing vector search technology to understand semantic nuances, GAUDS ensures that the search results are contextually tailored to the user's research intent. This process not only supports users in finding what they need but also helps them understand and use the data effectively, much like how a reference interview aims to empower the user with precise and useful information. This parallel enhances the GAUDS system's role in guiding researchers through their data journey, ensuring both discovery and effective reuse.

In general, GAUDS expedites decision making on the reuse of existing data by simplifying the evaluation of data availability and suitability, thus reducing the preliminary barriers to data reuse exploration in the *Initiation* stage. Moving into the *Exploration and Collection* stage, the system significantly enhances the capability to discover, acquire, understand and select relevant data. Incorporating interactivity and advanced data visualization tools allows researchers to dynamically navigate through data entities and contexts, effectively reducing the iterative cycles of data discovery and selection. This feature aligns perfectly with the need to refine choices and incorporate additional data to ensure the optimal fit for research purposes. In the *Repurposing* stage, GAUDS supports potential processing operations through data reuse examples in previous publications, allowing researchers to apply the proper methods to align the collected data with specific research needs. This facilitation minimizes the operational overhead associated with data management and enhances research efficiency, fostering an environment conducive to innovation. By mitigating the complexities inherent in each stage of data discovery and reuse, GAUDS ensures a more streamlined and efficient data-driven research process. Broadly speaking, the GAUDS system has great potential to assist and enhance general archival processing. Through proper digitization, any archives ranging from primary source documents, historical records, manuscripts, to other artifacts, can benefit from similar technological advancements. By applying GAUDS-like systems, these materials can become more accessible and usable for research, thereby preserving valuable historical and cultural information while facilitating new discoveries.

Notably, in the data reuse process, there are distinctions between discovery and reuse stages. During the discovery stage, the primary focus is on identifying and locating relevant datasets. This stage benefits significantly from high recall, ensuring that a broad array of potentially useful datasets are presented. Flexibility and relevance in search results are critical here, as researchers need to explore a wide range of data to identify what might be pertinent to their research questions. In contrast, the reuse stage emphasizes the practical application of the identified datasets. At this point, precision becomes more critical than recall. Researchers need to access detailed metadata and have interactive tools to customize and interpret the data to fit their specific needs. The GAUDS system's ability to provide direct access to detailed metadata and support extensive customizability is particularly valuable in this stage, as it allows researchers to tailor the data precisely to their research objectives.By understanding and addressing the distinct needs of these stages, the GAUDS system ensures that it supports the entire data lifecycle—from initial discovery to detailed reuse.

In conclusion, the GAUDS system represents a significant advancement in data search and utilization technologies. Its comprehensive approach to data handling—from initial query to dataset reuse—demonstrates a profound understanding of the needs of the research community. The transparent and responsible system not only supports individual research activities, but also contributes to the larger goal of fostering an integrated, innovative, and collaborative research landscape[11].

[11] Appendix B presents more examples of the use of GAUDS

Chapter 7

Analyzing Focus Group Findings of the GAUDS System

This chapter presents findings from focus groups of researchers using the GAUDS system. The study collected feedback on the system's functionality and user experience, aiming to understand how effectively it meets researchers' varied needs in the health and medical domain, which shed light on further enhancement of GAUDS. In particular, this user study focuses on the alpha version of the GAUDS system, with regard to its ability to provide relevant and high-quality data recommendations and understand its effectiveness in supporting research data needs. After a detailed introduction to the system through the "How-to Guide" and a search example, all participants independently experiment with the system. They tried at least three health and medical research questions and inputted them as natural language queries. When they encountered a technical problem, the participants asked the author for help. The author had not observed user activities while experimenting, but participants documented their tried queries and user experience in the in-session questionnaire. Based on the experiments, the participants then attended the discussion session and left their feedback interactively. We provide further details of the study in the Appendix C. Feedback gathered from the focus groups is instrumental in highlighting the strengths and areas of improvement of the system, ensuring that GAUDS continues to evolve.

The findings of the focus group study reveal significant optimism about the potential of the GAUDS system to transform the discovery of research data. The participants praised several aspects of the system that echoed the CEVI principles (Connectivity, Effectiveness, Visibility, and Interactivity) discussed in earlier chapters.

- **Connectivity**. Participants were particularly impressed with the GAUDS System's ability to visually and functionally connect diverse research entities, including datasets, publications, and authors. The system's graphical representations were noted for their clarity and detail, effectively illustrating complex research relationships. Suggestions for enhancing these graphs included increasing their size and adding interactive elements such as hyperlinks to nodes, which would facilitate deeper user engagement and a more intuitive navigation experience.

- **Effectiveness**. The system was lauded for its intelligent data aggregation and recommendation capabilities, which streamline the research process considerably. Participants noted that GAUDS could serve as a robust alternative to traditional search engines, particularly in academic settings where it supports in-depth project-based learning and research initiatives. The focus group appreciated the system's ability to quickly provide relevant data, although they also pointed out the necessity for continuous updates and verifications to ensure data relevance, particularly in fast-evolving and highly specialized fields like health and medicine.

- **Visibility**. While the system's design was generally well-received, participants identified several opportunities to enhance visibility and accessibility across different devices. Improvements such as optimizing the system for various screen sizes and ensuring compatibility with different user preferences, like dark mode, were suggested.
- **Interactivity**. The potential for a more interactive user experience was a recurrent theme in the feedback. Participants expressed a desire for more advanced data visualization tools and better organization of information, which would allow for easier navigation through large data sets.

7.1 Method: Focus Group Study

In this section, we describe the focus group study of the GAUDS system. To study the alpha version of the system, we choose the focus group study method since it makes sure the participants can directly use the system, interact with the moderator, and leave quality feedback based on their first-hand experience. Participant selection was strategic, involving six professionals[1] from distinct but relevant backgrounds who have research data needs in the health and medical domains, ensuring a comprehensive understanding of the system in various types of users. Table 7.1 shows the summary of their background, and we use their User IDs to refer to their answers in the later sections.

Table 7.1: Focus Group Participants' Background

ID	Background
1	Biostatistics Doctoral Student with a focus on clinical trial data analysis
2	Health Policy Research Associate with an interest in Health Management and Policy
3	Data Scientist with experience in data analysis, machine learning, and health informatics.
4	Medical Research Data Scientist working in a hospital
5	Information Science Master's Student with research experience and interests in health informatics
6	Biostatistics Master's Student with research experience and interests in bioinformatics

[1] These research professionals are or closely related to the author's collaborators, all of whom have or are pursing graduate degrees. Three (half) of them at least heard of ICPSR and data search. All are familiar with the health and medical domains.

The focus group study was scheduled for a total of 90 minutes in an in-person setting, where the author served as the moderator. The study design included a pre-session questionnaire to establish participants' backgrounds and expectations, a hands-on system interaction phase, and a structured discussion of user feedback.

The pre-session questionnaire gathered demographic information, prior experience with similar systems, and initial expectations about the GAUDS system. During the hands-on phase, participants engaged directly with the GAUDS system, experimenting with at least three dataset search tasks using the research questions each pariticpants are interested in (within the health and medical domains). For example, if a user is interested in finding a dataset to study ADHD, they may use the query ”What datasets are essential for ADHD research? Can you name the researchers involved?” Users were then asked to document their natural language queries and overall feedback in an in-session questionnaire. The participants then attended a group discussion about the features of the GAUDS system, where they provided detailed feedback on its functionality, the relevance of the search results, and the overall user experience.

We then identify themes and suggestions in the focus group study. The author recorded the full user study, documented key points in the discussions, and referred to the transcripts generated by the recording software, summarizing and culminating in a detailed report outlining key findings and recommendations for system enhancements. Below are the details of the questions in the focsu group discussion, including the general inquiry for the feedback of the system features and the structured discussion:

- **Features**:
  - *Most Useful Features*: Which features do you find most useful?
  - *Missing or Improvable Features*: Are there any features that are missing or could be improved?
- Structured Discussions:
  - **Connectivity**: *Dataset-Publication Landscape*: Are the information provided in the system sufficient for your understanding of the overall dataset-publication landscape?
  - **Efficiency**: *Application Effectiveness*: How effective is the application? *Workflow Integration*: How well does the application integrate with your usual workflow?
  - **Visibility**: *Visual Ease*: Does the system provide essential visual easiness to recognize research entities, i.e. the demonstration of their interrelations through visualization?
  - **Interactivity**: *Knowledge Synthesis and Application*: Can the system facilitate knowledge synthesis and application, with support for customizability and flexibilty in the search experience?
- **Privacy and Security**: *Data Privacy Concerns*: Do you have any concerns regarding data privacy and security? *Handling of Privacy and Security*: How do you believe the application handles this?

While the free inquiry was designed to obtain any user feedback and the structured inquiry was designed to elicit feedback directly related to the CEVI dimensions, all of the results, in fact, can be summarized into the CEVI dimensions. The analysis and quoted user response details are provided in the following sections, where each response was carefully analyzed and mapped into one or more of the CEVI dimensions. The mapping process is based on the qualitative analysis of the user feedback's main topic, especially the direct mentions of frontend, backend, or database components such as ”the sidebar”, ”the LLM's hallucination”, or ”the scope of the database”, aiming to both highlight strengths and pinpoint necessary improvements through the eyes of the GAUDS system users.

7.2 Themes from User Feedback

Based on the free discussion of features, the structured discussions around the CEVI dimensions, and the potential privacy and security concerns, we identified the corresponding themes of user insights. They highlight the utility of the current alpha version of the GAUDS system and provide recommendations for enhancements. Table 7.2 summarizes the overall themes from the focus group.

Table 7.2: Summary of Feedback Themes

Principle / Theme	Key Point
Connectivity	Related Research Suggestions: Participants highly valued the system's ability to suggest relevant research questions and datasets.
	Graphical Representation: Appreciation for clear illustrations of connections between datasets, authors, and publications.
	Clarity of Information Display: Praise for the clarity with which complex relationships are displayed, aiding decision-making of what datasets to use and how to use datasets.
Efficiency	System Responsiveness: Mixed reviews on the system's speed and responsiveness, especially with specific queries.
	Integration with Workflows: Discussion on how the GAUDS System could be integrated into existing research workflows.
	Specificity of Search Results: Challenges noted with precision in more specific queries.
Visibility	Mobile and Desktop Usability: Potential for mobile use noted, but need for better navigation aids on larger screens emphasized.
	Accessibility Issues: Difficulties with text visibility in different browser modes, suggesting broader accessibility needs.
	Clear Navigation Aids: Recommendations for easier to follow "How-to Guide" and clearly marked navigation examples such as video tutorials, especially for new users of the system.
Interactivity	Graphical Interface and Navigation: Suggestions for more intuitive design, which streamlines the user search path in a straightforward order based on the users' workflow, and better organization of information, which avoids long and hard to read outputs like JSON format.
	Data Presentation Enhancements: Calls for improvements in how data is presented, such as using histograms for numeric variables.
	User Interaction Tools: Desire for advanced tools allowing for more active data manipulation and understanding.

7.2.1 Mapping User Feedback to the CEVI Principles

Below, we discuss the themes that we observe in the user feedback. We provide direct quotes to further demonstrate the details in the focus group study and support our summary[2].

Connectivity. The GAUDS System's ability to effectively map the relationships between various research entities like datasets, publications, and authors was highly praised. Participants particularly valued features like the "Reuse" function, which suggests relevant secondary research questions and recommends datasets, thereby enhancing the potential research objects to use and comprehensively supporting the research scopes in their research projects. This connectivity of research objects related to datasets and publications is further evidenced by the system's comprehensive visual graphs, which clearly depict relationships such as shared authorship and journal sources. Feedback suggested that while the connectivity feature is robust, it could be improved by making the graphical representations larger and more interactive, potentially incorporating hyperlinks to research entities such as datasets, publications, and authors, and more straightforward system integration with existing resources such as additional metadata fields of these research entities on the ICPSR and OpenAlex websites.

Efficiency. Feedback on the GAUDS System's efficiency was varied. Many users found the system informative and straightforward, appreciating its design and the speed of precise search results. However, challenges were noted, particularly with the system's performance on more specific queries, which seemed to reduce in precision. The integration of the system into the users' existing workflows was viewed positively, suggesting that with additional resources like detailed user guides or video tutorials, the GAUDS System could become an integral part of their research process[3].

Visibility. While the system's layout was commended for potential mobile use, the feedback highlighted the need for clearer navigation aids, especially on larger screens. Issues with color contrast in different browser settings like dark mode were pointed out, indicating the necessity for a more adaptable design that accommodates a broad range of accessibility needs.

Interactivity. Participants found that the system's interactivity, particularly in terms of data manipulation and exploration, could be enhanced. Suggestions included better graphical interface designs to facilitate easier navigation of information and enhancements in data presentation, such as displaying histograms for numeric variables to aid in data understanding and manipulation.

In summary, the GAUDS System shows considerable promise in enhancing research efficiency through its connectivity and interactivity features. However, the insights from the usability testing emphasize the need for improvements in efficiency, visibility, and further development of interactive elements to enhance user engagement and satisfaction. These enhancements will likely increase the system's utility in the academic setting, fostering a more intuitive and effective research environment.

[2]We use *parentheses* to denote the author added words that make the spoken language clear and grammatically correct.

[3]While the users are not specific about their research process, we provide a summary of it in Chapter 6 Figure 6.12

7.2.2 Connectivity

A comprehensive knowledge graph-based system should ensure that all research entities and their underlying connections are well-captured. This theme summarizes our understanding of how well the GAUDS System captures the connectivity between different research entities such as datasets, publications, and authors. The users provided generally positive feedback about connectivity in GAUDS.

For example, Participant 3 valued the Reuse feature of the system, particularly how it suggests related research questions and relevant datasets, which could enhance their original research scope and assist in the literature review process.

> I especially like the reuse feature. I think one thing that I really liked about this feature is that it could suggest some other factors that's related to my original research question. For example, my first question about was about analyzing women's health. And one of the (research) questions that is suggested was about investigating whether the presence of children would impact women's health. So I think, this expands the scope of my original research and could potentially suggest other insights that I could incorporate into my research project. So I think that feature was very helpful.... And also suggested some data sets related to these secondary factors. That can also help my literature review. If I want to go that direction.
>
> I think it's pretty clear from the graph that it's provided. So from the graph, we can see, like, what publication does this data set connect to (and) what author is contributed to the publication.

Participant 6 commended the clarity of the system's graph, noting its effectiveness in displaying connections such as shared authorship or journal sources between datasets.

> The graph is excellent. You can clearly see how the two datasets are connected...whether they have the same author, or if they are from the same journal.

Overall, the feedback on the GAUDS System's ability to illustrate connectivity between research components was positive, highlighting the effective use of graphs. However, participants felt that these graphical representations could be enhanced by increasing their size and interactive capabilities, such as incorporating hyperlinks and better integration into the system's layout.

7.2.3 Effectiveness

A useful search application offers users a platform that effectively consolidates fragmented information, surpassing the capabilities of conventional search engines. We received feedback on both the *Overall Effectiveness* and the *Integration* with user's usual data search workflow.

Regarding the overall effectiveness, participants 2, 3, 4, and 6 express satisfaction with the GAUDS system, finding it very informative, easy to use, and particularly useful when additional variables are needed for an existing dataset.

> I feel like it's quite straightforward to use, because there isn't a lot of button everywhere. I actually found the search box very intuitive and picked up the system pretty fast...I feel like it can be effective when I already have a data set that I am working on, but I need some additional variables. (Participant 2)
>
> I really liked the GAUDS system; I think it was really informative. (Participant 3)
>
> Overall for me, it was very easy to use. (Participant 4)
>
> The GAUDS runs perfectly well. It gives search result very quick and are very precise for the question I asked. The interactive graph is really good. (Participant 6)

Participant 5 appreciated the detailed explanations provided by the GAUDS system, noting it offers a significant time-saving advantage over conventional search tools like Google. At the same time, this participant also pointed out the potential ”hallucination” since the system did not appropriately balance the weights of the different keywords in the query. The participant found that the system gave more weight to ”occupant” over ”visual acuity,” which led to search results that are more skewed towards datasets containing the term ”occupant,” neglecting the main interest in ”visual acuity.” This imbalance could be due to the internal ranking or frequency of terms within the datasets.

> I think with your explanation of how this dataset is gathered and the purpose of this (system), it is pretty good and could replace a lot of work from Google...for my second query, the occupancy part, there are tons of dataset(s) returned by this tool. And there are excellent explanations on why those data are related and how to use them. Key feature of this is it offers reasoning. But Google doesn't. Google just return the value so you really need to get in there and understand the logic, which might take a long time. This reasoning process is already done by this tool. So it is really time saving.
>
> I think there might be a little bit hallucination. I was interested in the relation between visual acuity and work occupancy. When I am using the Ask and Reuse, it returns more dataset regarding the word “occupant” instead of “visual acuity” related to “occupant”... it seems that the inputs do not weigh equally, and the inner collection of my inputs are not recognized.

While the word frequencies are all embedded in semantic vectors, this finding suggest the need for further investigation of different embeddings and different similarity metrics' usefulness in the GAUDS system's specific use case. Currently, the system is based on OpenAI's word embedding and the cosine similarity metric, as we discussed in Chapter 4. This feedback on the potential ”hallucination”, which was actually due to vector's imbalance of different necessary information, calls for further customization of the word vector and adaptation of similarity metrics in our system. Moreover, the system might lack a robust feedback mechanism to learn from user interactions and adjust its search algorithms accordingly. The GAUDS system can also benefit from implementing a multi-round conversion mechanism to mitigate the repeated issues in query handling and result relevance to further improve the user's experience.

Other participants also point out the current challenges of using the system. Participant 1 observed that the system's effectiveness diminishes with more specific queries, suggesting that while broad queries return relevant data, highly specific inquiries might result in less precise data retrieval. This lack of effectiveness on queires specific to scientific languages may due to three main reasons, including *Specificity and Relevance*, *Vector Representation Challenges*, and *Balance Between Precision and Generalization*.

> The more specific questions I ask, the less effective the returned dataset. For example, the first question I asked was broad: "If I am interested in a Phase I clinical trial for oncology, what kind of data can I use?" All the return data are about this question. My last question was more specific: "I am interested in Bayesian method in imaging data for Aortic Growth Mapping." And the return data are more like a broad interpretation of aortic growth, which is focused on cardiovascular disease. So maybe I need to change my question or rephrase it.

First, regarding Specificity and Relevance, the user's experience indicates a discrepancy in how the system handles broad versus specific queries. For broad queries, like those about Phase I clinical trials in oncology, the system returns highly relevant datasets because the query likely matches a wider array of indexed terms and metadata that are commonly present in the datasets. However, when queries become highly specific, such as asking about the Bayesian method in imaging data for Aortic Growth Mapping, the system struggles to return equally relevant results. This suggests a challenge in matching the specificity of user queries with appropriately detailed datasets, especially when the language use in the natural language query is highly specific to a research domain.

Second, for the Vector Representation Challenges, as noted, the potential issue might be related to how textual data is vectorized within the system. Vector representations, which are commonly used in natural language processing to handle and interpret text data, may not always capture the specificity needed in scientific contexts. Scientific terminology requires high precision, and vector spaces might dilute this by averaging or generalizing the semantic representation of terms. This is particularly problematic for specific queries where exact matches or close semantic relationships are crucial.

Third, the system appears to struggle to balance the need for precision in matching specific scientific terms and the broader capability to generalize from user queries. This balance is essential in scientific search systems, where both the broad context and the specific details can be crucial for retrieving relevant datasets. The GAUDS system might be overgeneralizing in its approach to handling specific queries, leading to less relevant results as it fails to capture and prioritize the detailed aspects of the query[4].

[4]The balance between specificity and sensitivity has long been a problem in data science. GAUDS, like any other search engine, is likely to provide irrelevant search results and thus need a comparatively high number of results returned to users. This prompted

In addition, participant 6 found the user interface confusing and suggested more specific user guidance.

> I think, to be honest, at first I was a little bit confused with the user interface. But I can understand that it has a very complex function for this (tool). So it is hard to design perfectly. Maybe the introduction (such as) the user guide can be more specific.

Regarding the effectiveness of potentially integrating the GAUDS system into user data search routines, participants 6 and 4 discussed the integration of the GAUDS system into their research process, suggesting enhancements such as step-by-step video tutorials to improve usability.

> It (GAUDS) fits in my experience pretty well. First, I will extract my research questions and query terms. Then I will go to some dataset providers such as Kaggle or Google. It (GAUDS) is another place that I can use to search for data. (Participant 6)
>
> Recording a step-by-step video tutorial would also be useful. (Participant 4)

Participant 3 highlighted the potential of the system to be integrated in educational settings, allowing students to explore and structure projects according to their research interests.

> I think this system is particularly useful in an educational setting. In schools, maybe sometimes this teacher will assign a specific project where the students can freely choose what topics they can delve into. So by using a system, they can search for the topic that they are interested in, but may not be that familiar. Then they can just learn by navigating through the system and then pick out the data set they want to use and then structure their project based on the data set they choose.

Overall, participants acknowledged the effectiveness of the GAUDS system in streamlining the research process by intelligently suggesting relevant information and datasets. The system's ability to act as a more efficient integrated alternative to traditional search engines like Google was noted as a significant advantage, particularly in academic and professional settings. The feedback also suggests that the GAUDS System might be beneficial in educational settings by helping to facilitate project-based learning. At the same time, search effectiveness may vanish when the query is specific with highly specialized terminologies, which can be improved through enhanced word vectors and user feedback loops.

---

another question: the readability (visibility) of the results. Thus, we experimented with different "Top K" values in Chapter 6 and found that a Top K of 30 helps the GAUDS system perform the best in recall. Similarly to search in Google and other search engines, users often focus more on what is "useful", as long as the returned results are not too many to read

7.2.4 Visibility

This dimension addressed the visibility and accessibility of the GAUDS System's features, especially in different user environments and devices.

Participant 5 praised the system's layout for mobile use but recommends clearer navigation aids on larger screens.

> I think the location (of the query bars) is pretty good, and also it could be mobile-friendly, if you are having a little screen, maybe you're using cell phone to do some research. So by having as a side (bar) is really helps to type my research question, and also manipulate the Reuse section. However, if I am on a regular laptop, since the query bar is on the side, there is no clear indication of where to input the query. So maybe a huge title bar would be helpful.

Participant 6 faced difficulty reading text in dark mode due to poor color contrast, suggesting a need for broader accessibility considerations.

> My browser is in dark mode, and then the text is also black. I can barely see the text on the graph... important to accommodate to broad accessibility in different settings

Participants identified pros and cons of visibility related to both the system's compatibility with different browser settings and its usability on various devices. The feedback points towards a need for a more adaptable and user-friendly design that ensures all users can efficiently interact with the system regardless of their device or settings.

7.2.5 Interactivity

Feedback in this dimension explored how the GAUDS System facilitates user interaction with data, specifically regarding data manipulation and exploration.

Participant 1 discusses difficulties interacting with the graph, such as viewing text on edges and navigating extensive information, suggesting improvements in graphical interface and information presentation.

> I can share more about the show Table & Graph section. When I type in over three questions, it's really hard for me to see clearly. I need to spread the nodes to see the text on the edges.
>
> Explore contains long information on one page. So it takes me time to scroll down up to the end to see what kind of results I have under each section. I'm guessing whether be more efficient to separate sections into different small windows. So we can click the button and see specific answer under the section call.

Participant 3 proposed enhancements in how data is presented to users, such as displaying column names and histograms for numeric variables, to facilitate better data manipulation and exploration.

> Maybe there is a way to present data to the users. For example, Kaggle presents the name of the columns and histograms of numeric variables. Would it be possible to include some features such as this? So you are not only suggesting the question to explore, but also data manipulation or exploration steps.

Despite the overall effectiveness discussed in the previous subsections, the feedback on interactivity emphasizes the need for more advanced data interaction tools within the GAUDS System that allow users to not only view but also manipulate and understand data more effectively. This could include better organization of information and enhanced data visualization tools.

7.3 Synthesized Future Development Plan

Overall, the results from the focus group show that users are interested in the GAUDS System as a new tool for finding and reusing research data. The feedback gives us a clear plan for future development, pointing out the need for better interactivity, visibility, and regular updates to the data sets to keep up with research progress. These findings will help us improve the GAUDS System, making sure it stays updated with technology advancements in data search systems and continues to support users' research work.

To inspire future research and practical enhancements, we further summarize the user suggestions into a comprehensive plan from the **Frontend**, the **Backend** (with algrithms and data flow), and the **Database** (with data source) perspectives. These enhancements aim to make the GAUDS system more user-friendly and responsive to the complex needs of researchers. By implementing these suggestions, the GAUDS System can enhance its usability as a research tool, making it a more valuable resource for the academic research community. Table 7.3 shows a summary of the suggestions for potential improvements.

Table 7.3: Proposed Enhancements for the GAUDS System

Suggestion	Key Point
Frontend	Enhance Interface Clarity: Improve the visibility and accessibility of key interface elements like the search box.
	Feature Explanations: Include clearer, more detailed descriptions or tutorials on the functionalities of different search modes.
	Enhance Graph Usability: Increase the size and visibility of the connectivity graphs, possibly by placing them in a dedicated tab that allows for interactive elements like zooming and scrolling.
	Integrate Hyperlinks: Add hyperlinks to the graph's nodes to provide quick access to detailed information on datasets, publications, and authors.
	Improve Responsive Design: Ensure the system is fully responsive and works seamlessly across all devices, including mobile phones and tablets.
	Adapt to User Settings: Integrate features that adapt to user-specific settings like dark mode to enhance accessibility and usability.
	Improve Information Organization: Organize information in a more user-friendly manner by allowing users to expand and collapse data sections as needed.
	Enhance Educational Features: Further develop features that support educational use, such as improved navigation and customized dataset recommendations based on the curriculum.
Backend	Error Handling Improvements: Develop more robust error handling mechanisms to prevent system crashes from syntax errors.
	Advanced Filtering Options: Introduce time-based filtering to prioritize newer datasets in search results.
	Cost Considerations: Evaluate the cost-effectiveness of features in relation to their utility and operational expenses.
	Improve Semantic Matching: Enhance the system's semantic analysis capabilities to ensure the relevance of suggested datasets and information.
	Implement Advanced Data Visualization Tools: Include features that allow users to visualize data distributions and other statistics directly within the system.
	Strengthen Data Access Controls: Implement robust mechanisms to control and monitor access to sensitive data.
	Enhance User Authentication Processes: Develop more stringent user authentication processes to prevent unauthorized access.
Database	Expand Dataset Collection: Continuously update and expand the database to cover a broader range of topics and ensure it includes the most current research data.
	Dynamic Publication Updates: Ensure the database includes the latest research to support fields with rapid developments.

Frontend. Improvements to the GAUDS System's frontend are centered on enhancing user interaction and data visualization. Key initiatives include boosting the visibility and accessibility of interface elements such as the search box and providing more comprehensive explanations or tutorials on the functionality of different search modes. There is also a significant emphasis on improving the usability of connectivity graphs by increasing their size and visibility, potentially situating them in a dedicated interactive tab that supports zooming and scrolling. This usability improvement will also fix the known bugs in the non-intuitive user interactions with the network graph, such as the zooming in of nodes after a single click. Further, integrating hyperlinks directly into the graph's nodes can provide quick access to detailed data on datasets, publications, and authors, fostering a more seamless research process.

As an end-user tool, ensuring the system is responsive across all devices, including mobile phones and tablets, is critical. Adaptations to accommodate user-specific settings, like dark mode, will enhance accessibility. Organizational improvements such as allowing users to expand or collapse data sections and enhancing features that support educational use, including better navigation and tailored dataset recommendations based on the curriculum, are proposed. Moreover, increasing interactive learning tools that enable direct data manipulation within the system can significantly enrich educational experiences.

Backend. Backend enhancements are focused on increasing the robustness and sophistication of the GAUDS System. There are several workflow-related aspects to be improved. First, there is a need to develop more comprehensive error handling mechanisms to prevent system disruptions caused by syntax errors is crucial. Second, considerations of the cost-effectiveness of features relative to their utility and operational expenses are also important. Third, strengthening data access controls and enhancing user authentication processes are essential to ensure ethical use of the tool and secure access to sensitive data.

For methodological aspects, enhancements in semantic matching are necessary to improve the relevance of dataset suggestions and information, especially for complex or specialized research topics. In addition, including advanced data visualization tools within the system will allow users to directly explore data distributions and other statistical analyses. Finally, introducing advanced filtering options, such as time-based filtering, can also help prioritize newer datasets in search results.

Regarding semantic matching models in the backend, in particular, there are several specific implementations that can help improve the system:

- Enhanced Vector Models: Implementing more sophisticated vector models that are better suited for scientific language could improve performance. Models that incorporate domain-specific knowledge and understand the context in which terms are used could provide more precise vector representations.

- Query Interpretation Enhancements: Enhancing the system's ability to interpret and prioritize elements of user queries can lead to better data retrieval. Techniques such as named entity recognition tailored to scientific contexts or query expansion mechanisms that consider the specific nuances of scientific research could be beneficial.

- Feedback Mechanism: Integrating a robust feedback system where users can indicate the relevance of returned results could help in fine-tuning the system's responses. This adaptive approach would allow the system to learn from specific instances and improve over time.

Database. For the database component, it is vital to continuously update and expand the dataset collection to encompass a broader range of topics and include the most current research data. Implementing dynamic publication updates is necessary to maintain a database that supports fields characterized by rapid developments, ensuring that researchers have access to the latest data to inform their studies.

Chapter 8

Conclusion and Future Directions

This dissertation has explored the potentials and limitations of integrating Generative AI, in particular Large Language Models (LLMs) and Retrieval Augmented Generation (RAG), with Scholarly Knowledge Graphs (SKGs) to enhance research data discoverability and reusibility. Through the development and analysis of the GAUDS system, this study has demonstrated significant advancements in data search technology, particularly in the health and medical domains. The findings of the focus groups and technical assessments provide a foundation for ongoing improvements and future work in the building of data search applications.

The remainder of this chapter concludes the dissertation by highlighting the significant contributions and outlining a clear path for future research and development. The integration of SKGs with Generative AI has not only demonstrated the potential to revolutionize data discovery and reuse, but has also set the stage for further innovations that could extend well beyond the academic and research data management sectors to user-centric search and recommendation systems.

8.1 Implications: From Technological Advancements in the GAUDS System to Data Search Stakeholders

The creation of SKGs from curated bibliographies in institutional repositories and less curated data sources in open databases represents a significant advancement in the field of information retrieval. Using Generative AI, particularly in the realm of LLMs and RAG, we can construct more dynamic and comprehensive SKGs. This approach not only reduces the reliance on labor-intensive curation processes but also enhances the richness and diversity of the knowledge graph. The implications for academic research are profound, where the ability to quickly assimilate and link diverse sources of information is crucial.

Direct user assistance in filtering information through Generative AI extends beyond mere data retrieval; it actively assists users in filtering information. In complex research scenarios, where users may be inundated with an overwhelming amount of data, Generative AI provides a nuanced and context-aware filtering mechanism. This enhancement is not limited to data alone, but extends to any form of digital information that users seek.

The use of LLMs as agents in information retrieval represents a paradigm shift in handling complicated search scenarios. LLMs, with their advanced natural language understanding capabilities, navigate through intricate queries to provide more accurate and contextually relevant results. This is particularly significant in academic and research settings, where queries often involve complex terminology and require deep domain knowledge. However, increasing delegation of research tasks to AI raises questions about the balance between technological assistance and the development of the researcher's own skills. It is essential to consider the potential trade-offs, such as the impact on the process of acquiring research expertise, and to examine how these tools might alter traditional research methodologies.

In light of the significant advances brought about by the integration of SKGs with Generative AI, it is crucial to consider the wide-ranging implications this technology has for various stakeholders. The GAUDS system not only enhances efficiency and interconnectivity, but also presents new challenges and opportunities for researchers, educators, students, funding agencies, data curators, and data managers. Each of these stakeholders plays a distinct role in the broader academic ecosystem, and their adaptation to these technological shifts will be pivotal in shaping the future of academic research and education.

For *researchers*, the integration of SKGs with Generative AI improves the ability to conduct interdisciplinary studies by quickly identifying relevant literature and data in various fields. This capability reduces the time spent on preliminary searches, allowing researchers to focus more on analysis and innovation. However, there is a potential risk that over-reliance on AI for information retrieval could impact the development of deep domain expertise, as researchers might lean more on AI tools than traditional research methods. Researchers must strike a balance, using AI to enhance productivity without undermining their critical analysis skills.

Educators can utilize these advanced information retrieval systems to develop more dynamic and current curricula that incorporate the latest research findings more swiftly. This technology can also be used to create more engaging and interactive learning experiences, where students can explore the interconnected aspects of knowledge through AI-driven visualizations of SKGs. However, educators must also ensure that students understand the foundational methods of research and critical thinking, beyond the AI-assisted retrieval processes.

Students stand to benefit significantly from direct access to enriched and dynamically updated knowledge graphs. Such tools can make the research process more efficient and accessible, particularly for those new to academic research. Additionally, the context-aware filtering provided by Generative AI can help students navigate through complex information landscapes more effectively, enhancing their learning and research capabilities. Nevertheless, there's a need for educational systems to instill traditional research skills to ensure students are not solely dependent on AI tools.

Funding agencies might see an opportunity in supporting projects that integrate SKGs with AI, as these projects could potentially lead to groundbreaking discoveries and improvements in information accessibility and analysis. Agencies might also consider the broader impact of such technologies on the research landscape, including the need for balanced funding to both technology-driven and traditional research methodologies. They might focus on initiatives that address the ethical implications and potential skill dilution in the research workforce due to AI dependency.

Data curators are impacted as the reliance on intensive manual curation is reduced, shifting their role towards overseeing the AI algorithms that assist in the curation process. This shift can allow curators to focus on higher-level tasks such as enhancing data quality, metadata standards, and the integrity of links within the SKGs. There is also an increased need for curators to possess a blend of domain expertise and technical skills to manage and oversee AI-driven curation processes effectively.

Data managers benefit from the ability to oversee more complex and expansive data sets with the assistance of AI-driven tools, which can automate the handling of repetitive tasks involved in data management. This technology can improve the accuracy and speed of integrating new data into existing SKGs. However, data managers must be vigilant about the potential issues related to data privacy, security, and ethical concerns associated with AI in data management. They also need to ensure that the systems are transparent and that the results produced by AI are interpretable and valid.

8.2 Limitations and Future Directions

As with any innovative technological endeavor, the development and deployment of the GAUDS system have inherent limitations that provide valuable learning opportunities and pave the way for future directions. This section discusses current limitations in the breadth, depth, and scope of the system and outlines planned expansions to address these challenges. Furthermore, we explore improvements needed in the underlying LLM-based intelligent systems that power the GAUDS, as well as the strategic development steps outlined in the system's roadmap. These discussions aim to set a clear path for scaling the system's capabilities and ensuring that it remains at the forefront of research data management technology.

8.2.1 Scope: Current Breadth, Depth, and Expansion Plan

The current scope of the GAUDS system, while effective as a prototype, is limited. Future expansion plans include incorporating more diverse data sources and focusing on the health and medical domains. These expansions involve the uniting of the main databases within the domain, such as those of the Chan Zuckerberg Initiative (CZI)[1] and institutional repositories beyond ICPSR. Another direction is federated data search and management [323], [324]. Scraping datasets' metadata from public information sources such as GitHub, academic papers, and research websites, this approach can suggest the dataset reuse and redirect to the owner's site for access. Moreover, the system can also connect to the entire research data lifecycle, such as Archival Data Repositories [325], to keep the search results relevant and up-to-date.

8.2.2 Models: Limitations and Improvements of LLMs-based Intelligent Systems

The LLMs themselves, while powerful, require continuous improvements to enhance their effectiveness in complex search scenarios. Future developments could include adopting new architectures such as Llama-3 [2] for larger, more complex models or Phi-3 [326] for smaller and more efficient models. Furthermore, more advanced LLM-based mechanisms in retrieval [327] and graph-based metadata filtering will be crucial [328].

[1]https://chanzuckerberg.com/

[2]https://ai.meta.com/blog/meta-llama-3/

8.2.3 System Development: Challenges and Roadmap

The development roadmap for the GAUDS system is described in Figure 8.1, which is presented below. This roadmap anticipates the iterative enhancements needed to address the evolving demands of research data management and discovery.

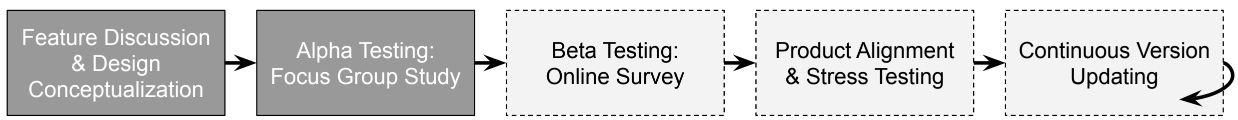

Figure 8.1: Development Roadmap of the GAUDS System

The development roadmap for the GAUDS system outlines a clear and structured approach towards refining and expanding the system's capabilities. The roadmap begins with the conceptualization phase, where initial ideas and features are brainstormed and designed. This phase sets the foundation for the entire project by establishing the goals and primary functionalities of the GAUDS system. Following the initial conceptualization, the system undergoes alpha testing through focus group studies. This phase is crucial because it allows for the collection of qualitative feedback directly from potential users, which is essential for identifying initial bugs and gathering suggestions for improvements. The insights gained here help refine the design and functionalities of the system.

The next phase involves beta testing, typically conducted through online surveys. This phase aims to reach a broader audience to validate the modifications made post-alpha testing and to ensure the system performs well under diverse usage scenarios. It provides a more quantitative measure of the system's performance and user satisfaction. Once beta testing confirms the robustness of the system and user approval, the roadmap includes product alignment and stress testing. During this stage, further adjustments are made to align the system more closely with user needs and industry standards. Stress testing ensures that the system can handle peak loads and complex queries without performance degradation. The final stage of the roadmap is continuous update of the version. In this ongoing phase, the system is regularly updated to incorporate new data sources, improve functionalities, and adapt to technological advances and user feedback. This continuous improvement cycle ensures that the system remains relevant and efficient in a rapidly evolving research landscape.

In general, this structured development process ensures that the GAUDS system is thoroughly tested and refined at each stage, guaranteeing a robust, user-friendly, and effective tool for the academic and research communities. Our findings and workflow also serve as a prototype for the development of other user-centric data search and recommendation systems, aiming to optimize user engagement and satisfaction across diverse informational landscapes.

Appendix A

Technical Details of the NPHardEval Benchmark

A.1 Examples of Synthesized Data, the Corresponding Prompts, and LLMs' Outputs

To further demonstrate the synthesized data, the corresponding prompts, and LLMs' outputs, we choose two specific problems with different attributes, including the EDP problem from the P complexity class with linear data synthesis and the GCP problem from the NP-Hard complexity class with graph data synthesis. We provide the zero-shot prompt for these questions and the output based on the GPT 4 Turbo. The details of all prompts and results are available at `https://github.com/casmlab/NPHardEval`.

A.1.1 Example for P: EDP

```
# Synthesized Data
{
    "string_a": "cef",
    "string_b": "ccb"
}

# Prompt with Synthesized Data
The Edit Distance Problem (EDP) involves finding the minimum number of operations required to transform one string into another, where each operation is either an insertion, deletion, or substitution of a single character.
Find the minimum number of operations required to transform the first string cef into the second string ccb. The operations are insertion, deletion, and substitution of a single character, each requiring 1 edit operation.
Please provide the minimum number of operations required to transform the first string into the second string. Provide a brief explanation of your methodology. Keep your response concise and focused.
Enclose your output within <root></root> tags. Present your reasoning in <reasoning></reasoning> tags and the final minimum number of operations in <final_answer></final_answer> tags, like <final_answer>{'Operations': 'MINIMUM_NUMBER_OF_OPERATIONS'}</final_answer>.
```

```

# Output (formatted for easier reading)
<root>
    <reasoning>
        To transform \"cef\" into \"ccb\", we can follow these steps:\n
        1. Substitute 'e' with 'c' (cef -> ccf)\n
        2. Substitute 'f' with 'b' (ccf -> ccb)\n
        This results in two substitution operations.
    </reasoning>
    <final_answer>
        {"Operations": 2}
    </final_answer>
</root>
```

Listing A.1: EDP Example

A.1.2 Example for NP-Hard: GCP

```
# Synthesized Data
c This is a generated graph with chromatic number 2
p edge 6 4
e 1 6
e 2 6
e 3 4
e 3 5

# Prompt with Synthesized Data
Graph Coloring refers to the problem of coloring vertices of a graph in such a way that
     no two adjacent vertices have the same color.
There are 6 vertices 1 to 6 in a graph. You may use 4 colors with alphabats from A, B,
    C,... to color the graph.
Label every vertex, even if it is disconnected from the rest of the graph. Please
    provide each vertex's color. Do not skip any vertices. You should also provide very
     short step-by-step reasoning. Do not use multiple lines and try your best to save
    output tokens.
Your output should contain two parts enclosed by <root></root>. First, your step by
    step reasoning wraped by <reasoning></reasoning>. Second, the final output of all
    vertex numbers and their associated colors, wrapped by the final_answer tag, like <
    final_answer>{0:'COLOR_1', 1:'COLOR_2', ...}</final_answer>.
The graph is as follows: Vertex 1 is connected to vertex 6.
    Vertex 2 is connected to vertex 6.
    Vertex 3 is connected to vertex 4.
    Vertex 3 is connected to vertex 5.

# Output (formatted for easier reading)
<root>
    <reasoning>
```

```
        Start with vertex 1, color it A; color adjacent vertex 3 with B; vertex 2 can be A as it is not adjacent to 1; vertex 4 connected to 2 must be different, use B; vertex 5 connected to both 2 and 3, use C; vertex 6 connected to 3 and 4, use A.
    </reasoning>
    <final_answer>
        {1:'A', 2:'A', 3:'B', 4:'B', 5:'C', 6:'A'}
    </final_answer>
</root>
```

Listing A.2: GCP Example

A.2 Details of Complexity Classes

There are nine category problems (tasks) in total in our benchmark, and each complexity class has three unique problem categories.

P (Polynomial time) Tasks

This class consists of tasks that can be solved by a deterministic Turing machine in polynomial time. Essentially, it represents tasks that are efficiently solvable. We include three P problems in the benchmark, namely Sorted Array Search (SAS), Edit Distance Problem (EDP), and Shortest Path Problem (SPP).

Sorted Array Search (SAS) SAS is about finding the position of a target value after sorting a given array. Given an array A of n elements and a target value T, the goal is to determine the index at which T is located in A after sorting. Renowned algorithms like binary search efficiently accomplish this task by iteratively halving the search interval, operating in logarithmic time. The problem can be formally stated as finding an index i such that $A[i] = T$, or determining that no such index exists. It is commonly used in databases and search engines to quickly find specific data within a large dataset [329].

Edit Distance Problem (EDP) EDP is about finding the minimum number of operations required to transform one string into another. Given two strings, A and B, of lengths m and n, respectively, the aim is to determine the minimum number of operations needed to convert A into B. The allowable operations are insertion, deletion, and substitution of a single character. Formally, the problem can be defined as finding a minimum number d such that the string A can be transformed into the string B using operations d. This algorithm has a time complexity of $\mathcal{O}(ab)$ where a and b are the lengths of the strings. When the complete dynamic programming table is constructed, its space complexity is also $\mathcal{O}(ab)$. EDP has widespread applications, especially in fields like computational biology for sequence alignment, natural language processing for spell checking and correction, and data analysis for measuring similarity between data strings.

Shortest Path Problem (SPP) SPP is about finding the shortest path between two nodes in a non-negative weighted graph. In our experiments, we ask for the shortest path between the first and last nodes. Given a graph $G = (V, E)$ with a weight function $w : E \to \mathbb{R}$ assigning weights to edges, and two vertices u and v in V, the task is to find the path from u to v that minimizes the total weight. This is often solved using Dijkstra's algorithm, which systematically expands the shortest path from the starting node until it reaches the target node. Formally, the problem is to find a path $P = (v_1, v_2, ..., v_k)$, where $v_1 = u$ and $v_k = v$, such that the sum of the weights of the consecutive edges in P, $\sum_{i=1}^{k-1} w(v_i, v_{i+1})$ is minimized. This problem can be used in network routing, GPS navigation systems, and logistics to find the shortest or most efficient path between two points. It helps reduce travel time and costs in transportation and communication networks.

NP-complete problems

This is a subset of NP. A problem is NP-complete if it is in NP and as hard as any problem in NP. If any NP-complete problem can be solved in polynomial time, then every problem in NP can also be solved in polynomial time. We include three NP-complete problems that are not in P in the benchmark, namely Traveling Salesman Problem Decision Version (TSP-D), Graph Coloring Problem Decision Version (GCP-D) and Knapsack Problem (KSP).

Traveling Salesman Problem (Decision Version, TSP-D) TSP-D is concerned with determining whether a salesman can complete a route, visiting each city at least once, with the total travel distance less than a specified value. Given a complete graph $G = (V, E)$ with vertices V representing cities and edges E representing paths between cities, each edge (i, j) is assigned a distance $d(i, j)$. The decision version of this problem asks whether there exists a tour (a sequence of cities) such that the total distance of the tour is less than or equal to a given value D. Formally, the problem can be stated as finding a permutation P of the set of cities $1, 2, ..., n$ that satisfies the condition $\sum_{i=1}^{n-1} d(P(i), P(i+1)) + d(P(n), P(1)) \leq D$. This problem is useful in logistics and supply chain management in planning efficient delivery routes and schedules [330].

Graph Coloring Problem (Decision Version, GCP-D) GCP-D involves determining whether it is possible to color the vertices of a graph using a given number of colors so that no two adjacent vertices share the same color. Given an undirected graph $G = (V, E)$, with V representing vertices and E representing edges, the goal is to find out if there is a way to assign one of k colors to each vertex such that for any edge $(u, v) \in E$, the vertices u and v have different colors. The formal statement is to determine if there exists a coloring function $c : V \to 1, 2, ..., k$ such that for every edge $(u, v) \in E$, $c(u) \neq c(v)$. It has wide applications in Round-Robin Sports Scheduling, Aircraft Scheduling, and Biprocessor tasks [331].

Knapsack Problem (KSP) KSP asks whether a subset of items can be chosen to fit into a fixed capacity knapsack without exceeding it, while also maximizing the total value of the selected items. Consider a set of items, each with a weight w_i and a value v_i, and a knapsack with a weight capacity W. The problem is to select a subset of these items such that the total weight does not exceed W and the total value is maximized. Formally, let x_i be a binary variable that indicates whether the item i is included in the knapsack ($x_i = 1$) or not

($x_i = 0$). The problem can be stated as the maximization $\sum_{i=1}^{n} v_i x_i$ subject to the constraint $\sum_{i=1}^{n} w_i x_i \leq W$, where n is the number of items. It is used in resource allocation and budgeting where the goal is to maximize the total value of a selection under a weight or cost constraint. Applications include cargo loading and electric vehicle charging [332], [333].

NP-hard problems

These problems are at least as hard as the hardest problems in NP. They may not necessarily be in NP (i.e., they may not have solutions verifiable in polynomial time), but solving an NP-hard problem in polynomial time would imply that P = NP. We include three NP-hard problems that are not reducible to NP-complete problems in the benchmark, namely Traveling Salesman Problem Optimization Version (TSP), Graph Coloring Problem Optimization Version (GCP), and Meeting Scheduling Problem (MSP).

Traveling Salesman Problem (Optimization Version, TSP) TSP-O involves finding the shortest route for a salesman to visit each city exactly once and return to the starting city. Given a complete graph K_n with n vertices, where each vertex represents a city and each edge (i, j) is assigned a non-negative cost or distance $d(i, j)$, the problem is to find the shortest possible route that visits each city exactly once and returns to the origin city. Formally, let P be a permutation of the set of cities $1, 2, ..., n$ representing the order in which the cities are visited. The travel salesman problem can be formulated as finding the permutation P that minimizes the total cost of travel, given by the function $f(P) = d(P(n), P(1)) + \sum_{i=1}^{n-1} d(P(i), P(i+1))$. This problem is important in operational research and logistics to find the most efficient route to visit multiple locations and return to the source, particularly route planning for delivery services, maintenance operations, and sales.

Graph Coloring Problem (Optimization Version, GCP) GCP-O refers to the problem of coloring vertices of a graph in such a way that no two adjacent vertices have the same color. Given an undirected graph $G = (V, E)$, where V is the set of vertices and E is the set of edges, assign a color to each vertex such that no two adjacent vertices have the same color. Formally, let $c : V \to C$ be a function that assigns a color from a set of colors C to each vertex in V. The graph coloring problem can be formulated as finding a proper coloring, i.e. a function c such that for every edge $(u, v) \in E$, $c(u) \neq c(v)$. This problem is used in constraints satisfaction problems and is applied in exam timetabling and register allocation in compilers [334].

Meeting Scheduling Problem (MSP) MSP deals with allocating time slots for meetings such that all constraints, including participant availability and room capacity, are met without overlaps. Given a set of n participants and their availability for m time slots, find a schedule that maximizes the number of participants who can attend the meeting. Formally, let $A = a_1, a_2, ..., a_n$ be the set of participants and $T = t_1, t_2, ..., t_m$ be the set of time slots. For each participant a_i, let S_i be a subset of T representing the times when a_i is available, and m_i be a subset of meetings that are required to attend. The meeting scheduling problem can be formulated as finding a subset $S \subseteq T$ such that $|a_i \in A | S_i \cap S \neq \emptyset|$ is maximized. In other words, the objective is to find a scheduling subset S_i where the collective availability of participants intersects with S_i, ensuring maximum participation. This problem is crucial in organizational management to schedule meetings that involve multiple participants with varying availability. It ensures optimal utilization of time and resources and is used in corporate scheduling systems and collaborative software [335].

A.3 Choices of Problems

In the benchmark, we exclude calculation-only (math intensive) tasks for each of the complexity classes, due to the overlap with already existing benchmarks and the known uncertainty of LLMs' math ability. For other reasoning, we provide detailed explanations and highlight them in bold.

A.3.1 Excluded P problems

Prime Number Determination Using algorithms like the AKS primality test to determine if a given number is prime. Reason: Math-intensive.

Solving Linear Equations Finding solutions for a system of linear equations. Reason: Math-intensive.

Maximum Flow Problem Finding the maximum flow from a source node to a sink node in a flow network. A flow network is a directed graph $G = (V, E)$ where each edge $(u, v) \in E$ has a capacity $c(u, v)$ and flow $f(u, v)$, with a designated source s and sink t. The objective is to maximize the total flow from s to t under the constraints that the flow on an edge does not exceed its capacity and that the incoming flow is equal to the outgoing flow for every vertex except s and t. Reason: **Most open source algorithms cannot follow the question and the prompt to provide outputs with mostly correct formats**.

A.3.2 Excluded NP-Complete problems

3-SAT Problem Deciding whether a given Boolean formula in conjunctive normal form with three literals per clause is satisfiable. Reason: Math-intensive.

A.3.3 Excluded NP-hard problems

Integer Linear Programming Finding the best integer solution for a set of linear equations and inequalities. Reason: Math-intensive.

Appendix B

Additional Use Cases and Feedback of the GAUDS System

B.1 Experimented Queries by Users

In this section, we provide all 18 participant queries experimented in the focus group study in Chapter 6. For further references, we call them Participant Query (PQ) and numbers using participant ID and the query ID. For example, the first PQ for participant 1 is numbered PQ1.1. If there are data sets related to participant-selected data sets, we also document them after each PQ. We also add notes associated to each PQ if the participant provides them in-text.

```
If I am interested in a Phase I clinical trial for oncology, what kind of data can I
    use?
```

Listing B.1: PQ 1.1

ID: 2539; 2540; 2562; 8491.

```
If I am interested in Alzheimer's disease research, what kind of data can I use?
```

Listing B.2: PQ 1.2

ID: 36053; 36036; 9915; 7669.

```
I am interested in the Bayesian method in imaging data for Aortic Growth Mapping. What
    kind of data can I use?
```

Listing B.3: PQ 1.3

ID: 9206; 36985.

```
What datasets are available for studying dementia in populations over 65 years old?
```

Listing B.4: PQ 2.1

ID: 2877; 3417; 36589.

```
What dataset I can use regarding to depression for youth?
```

Listing B.5: PQ 2.2

ID: 9973; 20240; 22121.

```
I would like to do a research on lung cancer among Eastern Country, what dataset should
     I use?
```

Listing B.6: PQ 2.3

ID: 121; 128; 154.

```
I am interested in analyzing womens health. what data set should I use?
```

Listing B.7: PQ 3.1

```
What datasets are available that detail the incidence and prevalence rates of Type 2
    diabetes in different age groups across various geographical regions? What datasets
     are available for studying Type 2 diabetes?
```

Listing B.8: PQ 3.2

The participant thinks that the first query in PQ 3.2 is too specific, so they added the second.

```
What databases contain information for oncology drugs?
```

Listing B.9: PQ 3.3

```
I plan to analyze the trends in tuberculosis. What are the recommended datasets?
```

Listing B.10: PQ 4.1

ID: 121; 128.

```
What datasets are essential for ADHD research? Can you name the researchers involved?
```

Listing B.11: PQ 4.2

ID: 20240; 9088.

```
Which are the leading datasets used in studying cervical cancer?
```

Listing B.12: PQ 4.3

ID: 2958; 8469.

```
How does long term exposure to allergy sources could affect the overall health
    situation? How would I gain data for this topic?
```

Listing B.13: PQ 5.1

```
I am interested in understanding long term effects of using Semaglutide and how it
    would affect the long term health condition of the user.
```

Listing B.14: PQ 5.2

```
I am interested in the relation between visual acuity and work occupancy.
```

Listing B.15: PQ 5.3

```
I'm focusing on genetic mutations in breast cancer. Which datasets include genetic
    profiles?
```

Listing B.16: PQ 6.1

ID: 2090; 4368.

```
I am interested in studying the cell perturbation cellular responses. What dataset
    should I use?
```

Listing B.17: PQ 6.2

ID: 2359; 2540.

```
Can you list the most cited datasets that explore the relationship between genetics and
     skin cancer outcomes?
```

Listing B.18: PQ 6.3

ID: 8826; 2090.

B.2 User Feedback

B.2.1 Expectations

GAUDS System Expectations: What are your current challenges in health and medical data discovery? What features do you hope to see in the GAUDS System?

Participant 1

> Challenges: There's a lack of information regarding the dataset's limitations. When the dataset spans an extended time period, data quality tends to be low. Expectation: We expect to see a description of the data quality and limitations in the GAUDS System.

Participant 2

> Hard to find datasets with information I need for my research question. Hope to see list of datasets (including variables provided) related to my research question using the system.

Participant 3

> Challenges: unable to find most suitable datasets for project; some time the codebook for specific columns and data attributes are missing; restricted access. I hope that the GAUDS System can allow for

> intuitive data finding, and contain other relevant information about the dataset and suggest how to use them.

Participant 4

> The most challenges for me is that where to start with my research, how can I find the dataset that I need for my research. The features that I hope to see in the GAUDS is that the database can be updated monthly to have more dataset.

Participant 5

> Extended time to look for data set.

Participant 6

> Hard to find health-related dataset to do biostatistical research.

B.2.2 Encountered Issues

Describe any issues you encountered while using the GAUDS System.

Participant 1

> I like the ”Show Table & Graph” section, but when I input too many datasets, the graph looks crowded and less clear. Additionally, when I use very specific search keywords, the output datasets may not effectively match the questions. Interestingly, after I input the last query, the returned datasets were more closely aligned with my second query – the datasets were mentioned under section: Results based on query vector’s similarity to datasets.

Participant 2

> No.

Participant 3

> I entered a queries that was too specific: ”what datasets are available that detail the incidence and prevalence rates of Type 2 diabetes in different age groups across various geographical regions?” no

> results was obtained using ”Ask” from the system.

Participant 4

> When the question is a bit similar, it will output the same data.

Participant 5

> The data set does not contain the most recent studies, e.g., Semaglutide. The the reuse section has a relative strict parsing rule (extra ; will cause crash).

Participant 6

> The GAUDS runs perfectly well. It gives search result very quick and are very precise for the question I asked. The interactive graph is really good.

B.2.3 Suggested Improvements

Suggest anything that needs to be improved.

Participant 1

> N/A.

Participant 2

> Give options to rank datasets by time.

Participant 3

> Sometime the top results are not the most relevant. For example, searching for type 2 diabetes, the first dataset returned was on mental health and depression – id: 201.

Participant 4

> Maybe can provide a little bit more information on the dataset list rather than just the topic name.

Participant 5

> It seems that the ai model could not understand the logic relation between the term with in the input. The query is ”I am interested in the relation between visual acuity and work occupancy”. The return data id_icpsr is 36054. The reasoning is ”How does visual acuity differ between young and older adults across the visual field?”. However, this reasoning deviate/not related to the main research question.
>
> It seems that data set could be further enriched. query : I am interested in understanding long term effects of using Semaglutide and how it would affect the long term health condition of the user. Since it is new medicine, no current related data is in the db.
>
> The the search box could be more obvious.
>
> More explanation regrading ””ask”” and ””explore”” could be added.
>
> It seems that the reasoning in reuse part is not entirely correct.

Participant 6

> The UI is a little bit confusing if didn’t look through the guide carefully.

B.3 The Data Documentation Details of ”National Health Interview Survey, 1994: Year 2000 Objectives Supplement”

The following example shows that the reuse suggestions are based on detailed data documentations. However, the results are usually too long to present in a single web page, which could lead to visibility issues. Thus, the system can improve by incorporating a more simplified and streamlined data documentation results, likely in a table, to present on the website. The full record can then be stored in a JSON file and made downloadable to the users.


```
"Owner": [
{
  "owners": {
    "name": "NACDA"
  }
}
],

"Funder": [],

"Most cited publications": [
{
  "p": {
    "year": 2000,
    "name": "Mobility impairments and use of screening and preventive services",
    "id_openalex": "https://openalex.org/W1604876104",
```

```
      "related_works": [list omitted for simplicity],
      "datasets": 6875,
      "reference_count": 339
    },
    "citations": 1
  },
  {
    "p": {
      "year": 2006,
      "name": "Occupation and lung cancer mortality in a nationally representative U.S. cohort: The National Health Interview Survey (NHIS)",
      "id_openalex": "https://openalex.org/W1978488746",
      "related_works": [list omitted for simplicity],
      "datasets": 6724,
      "reference_count": 13
    },
    "citations": 1
  },
  {
    "p": {
      "year": 2004,
      "name": "Trends in U.S. smoking rates in occupational groups: The National Health Interview Survey 1987-1994",
      "id_openalex": "https://openalex.org/W2040808418",
      "related_works": [list omitted for simplicity],
      "datasets": 9914,
      "reference_count": 101
    },
    "citations": 1
  },
  {
    "p": {
      "year": 2007,
      "name": "Smoking rate trends in U.S. occupational groups: The 1987 to 2004 National Health Interview Survey",
      "id_openalex": "https://openalex.org/W2070045664",
      "related_works": [list omitted for simplicity],
      "datasets": 4349,
      "reference_count": 100
    },
    "citations": 1
  },
  {
    "p": {
      "year": 2000,
      "name": "Firearm storage patterns in US homes with children",
      "id_openalex": "https://openalex.org/W1503382787",
      "related_works": [list omitted for simplicity],
```

```
      "datasets": 6875,
      "reference_count": 147
    },
    "citations": 1
}
],

"Most frequent author": [
{
    "a": {
      "author_id": "https://openalex.org/A5051038042"
    },
    "publications": 3
},
{
    "a": {
      "author_id": "https://openalex.org/A5054516354"
    },
    "publications": 3
},
{
    "a": {
      "name": "Sarah J. Newman",
      "author_id": "https://openalex.org/A5081055071"
    },
    "publications": 3
},
{
    "a": {
      "author_id": "https://openalex.org/A5073284312"
    },
    "publications": 3
},
{
    "a": {
      "author_id": "https://openalex.org/A5061866824"
    },
    "publications": 2
}
],

"Related datasets": [
{
    "d2.id_icpsr": 6344,
    "d2.name": "National Health Interview Survey, 1992: Cancer Control Supplement",
    "citations": 5
},
{
```

```
  "d2.id_icpsr": 9914,
  "d2.name": "National Health Interview Survey, 1990: Health Promotion and Disease
    Prevention (HPDP) Sample Person Supplement",
  "citations": 5
},
{
  "d2.id_icpsr": 6053,
  "d2.name": "National Health Interview Survey, 1991: Health Promotion and Disease
    Prevention Supplement",
  "citations": 5
},
{
  "d2.id_icpsr": 6724,
  "d2.name": "National Health Interview Survey, 1994",
  "citations": 4
},
{
  "d2.id_icpsr": 6343,
  "d2.name": "National Health Interview Survey, 1992",
  "citations": 3
}
],

"Restriction": [
{
  "d.restriction": "AVAILABLE"
}
]
```


Listing B.19: Data Documentation Example Output

B.4 Known Issues

In addition to the mentioned known issues described in Chapter 4 , Chapter 6, and Chapter 7, there are some additional knwon issues:

Author ID Due to a known recent change in OpenAlex's author ID, some author name and IDs may mismatch. We will fix the mismatching author information in the next version of the GAUDS system[1].

Relation Name The authored_by relation should be renamed to used_by, which we will correct in the next version of GAUDS. These relation names are temporary and may be different from the IHSKG's version. Thus, we will also consider simply remove them, which can also make the graphs clearer in presentation.

[1]For the ID update details, please see `https://groups.google.com/g/openalex-users/c/rDA7PWTarVQ`.

Click Effect The single click on nodes in the Show Table & Graph feature will make the graph ”dissapear” due to an unexpected zoom out. The user can either scroll down to zoom in or click on the Show Table & Graph button again to regenerate the graph. Ideally, through a single clicking, the system can open a new web window for the link of the node entity.

Appendix C

Supplemental Details of the Focus Group Study

This focus group study is approved by the IRB HSBS at the University of Michigan under EXEMPTION 2(i) and/or 2(ii) at 45 CFR 46.104(d) (UM Federalwide Assurance: FWA00004969). We provide more details of the focus group study as follows.

C.1 Formal Objectives Definition

The primary objective of this focus group study is to collect data on the usability, features and sanity of the GAUDS System, our proposed research data discovery and reuse application. Specific goals include:

- Assessing the GAUDS System's ability to meet the research needs of various health and medical informatics domains.
- Evaluating the relevance and quality of the data search and recommendation results.
- Identifying user experience improvements in the GAUDS system, especially compared to the current ICPSR Find Data web application.

C.2 Logistics and Preparations

C.2.1 Overall process

- Format: In person.
- Duration: Schedule the session for 90 minutes to allow a thorough discussion without causing fatigue.
  - 10-20 minutes: Participants fill out the presession questionnaire.
  - 20-50 minutes: Usability testing where participants execute their prepared queries.
  - 50-80 minutes: Structured discussion led by the moderator using the guide below.
- Recording: Video record the session for analysis, ensuring that all participants consent.

- Moderation: The author serves as the moderator since he is experienced in conducting a focused group study and knowledgeable about the application and domain.

C.2.2 Actionable items to prepare

- Pre-session Questionnaire: Prepare a brief questionnaire for participants to fill out beforehand, capturing their background, previous experience with similar applications, and specific interests in health data.
- Discussion Guide: Create a structured discussion guide with open-ended questions and prompts to facilitate conversation.
- During the focus group study, each participant needs to try at least three natural language queries (and document them in the In-session Questionnaire). Based on those, please participate in the discussion and leave a review of the system (also in the In-session Questionnaire).

C.3 Questionnaires

The questionnaires are facilitated through a Google form. Participants are asked to complete each part of the questionnaire in the corresponding period before the focus group study session. There are 13 questions in total. Most questions that expect a quantitative answer simply serve as a prediscussion that familiarize the participants with the study; the answers to the qualitative questions, as well as the focus group discussing after filling the questionnaires, are the focuses of this study design. The questionnaires are included on the following pages.

A Focus Group Study for GAUDS

Thanks for participating in the focus group study of GAUDS, a generative AI-augmented and user-centric data search (GAUDS) system. In this system, you can input a natural language query to search for a dataset you want to use for your research.

Please find the link to the website in your email.

This questionnaire has the pre-session and the in-session parts. The **pre-session part** is designed to capture your background information, previous experience with similar data search applications, and specific interests in health and medical data. The **in-session part** is for you to document experience during the session, focusing on usability tests and discussion contributions.

* Indicates required question

1. Email *

____

Pre-section Questionare

The **pre-session part** is designed to capture your background information, previous experience with similar data search applications, and specific interests in health and medical data.

2. Fist and Last Name (won't be shared with anyone)

____

3. Background (Please specify your current role and main area of expertise)

4. Prior experience with data discovery applications

Mark only one oval.

- Extensive Experience
- Some Experience
- No Experience

5. GAUDS System Expectations: What are your current challenges in health and medical data discovery? What features do you hope to see in the GAUDS System? *

6. Do you consent to the recording, transcribing, and analysis of the focus group discussion, as well as this questionnaire?

Mark only one oval.

- Yes
- No

In-session Questionare

The **in-session part** is for you to document experience during the session, focusing on usability tests and discussion contributions.

7. Usability Test Log: Document at least three natural language queries you performed and the corresponding dataset IDs you chose. *

8. For the three queries, overall, do you think the results (including the recommended datasets and the reuse guide) you obtained from the system are useful? *

   Mark only one oval.

   - Very useful
   - Somehow useful
   - Not useful

9. Which features (button) do you find most useful? Choose as many as you like.

   Check all that apply.

   - [ ] Explore
   - [ ] Ask
   - [ ] Find
   - [ ] Reuse
   - [ ] Show Table & Graph

10. Describe any issues you encountered while using the GAUDS System. *

11. Suggest anything that needs to be improved.

12. After participating in the discussion, how would you rate the GAUDS System on the following aspects? - Connectivity: Ensuring that all research entities and their underlying connections are well-captured, providing a comprehensive view of the dataset-publication landscape.
- Effectiveness: Offering users a platform that effectively consolidates fragmented information, surpassing the capabilities of conventional search engines.
- Usability: How intuitive is the application's interface? Are there any challenges in navigating the application?
- Integration: How well does the application integrate with your usual workflow?
- Visibility: Enhancing the users' visual easiness to recognize research entities, their interrelations through frontend visualization tools.
- Interactivity: Facilitating knowledge synthesis and application, with the agent offering support for interactive, customizable, and flexible queries.

Mark only one oval per row.

	1 - Poor	2	3	4	5 - Excellent
Connectivity					
Effectiveness					
Usability:					
Integration					
Visibility					
Interactivity					

13. Privacy and Security: Do you have any concerns regarding data privacy and security? How do you believe the application handles this? *

Appendix D

Papers and Datasets in Dissertation

In this appendix, we map the papers and datasets used in this dissertation to the manuscripts that are published, accepted, or in preparation. Table **??** shows the detailed mapping between chapters and scholarly work.

Table D.1: Mapping between Chapters and Scholarly Work

Item and Citation	**Chapter**	**Notes**
Field Preliminary (Unpublished)	2	In preparation to submit to JASIST (Original Title: Discoverability and usability enhancement in research data management: From the FAIR principles to scholarly knowledge graph implementations)
DataChat (ASIS&T) [290]	3	In preparation to republish in JASIST (together with Field Preliminary and SimSearch)
SimSearch (Unpublished)	3	In preparation to submit to JASIST (Original Title: S^3R^3: A scholarly semantic similarity benchmark for research relevance ranking between datasets and publications)
ICPSR Dataset [257]	4	Dataset before transformation to IHSKG (Original Title: A dataset for measuring the impact of research data and their curation)
NPHardEval [265]	5	Accepted to ACL 2024 (Original Title: Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes)
GADUS System (Unpublished)	6 & 7	In preparation to submit to JAMIA

Bibliography

[1] D. J. Conrado, M. O. Karlsson, K. Romero, C. Sarr, and J. J. Wilkins, "Open innovation: Towards sharing of data, models and workflows," *European Journal of Pharmaceutical Sciences*, vol. 109, S65–S71, 2017.

[2] J. C. Wallis, M. S. Mayernik, C. L. Borgman, and A. Pepe, "Digital libraries for scientific data discovery and reuse: From vision to practical reality," in *Proceedings of the 10th annual joint conference on Digital libraries*, 2010, pp. 333–340.

[3] N. A. Vasilevsky, J. Minnier, M. A. Haendel, and R. E. Champieux, "Reproducible and reusable research: Are journal data sharing policies meeting the mark?" *PeerJ*, vol. 5, e3208, 2017.

[4] S. Samuel and B. König-Ries, "Understanding experiments and research practices for reproducibility: An exploratory study," *PeerJ*, vol. 9, e11140, 2021.

[5] A.-W. Harzing and S. Alakangas, "Google scholar, scopus and the web of science: A longitudinal and cross-disciplinary comparison," *Scientometrics*, vol. 106, pp. 787–804, 2016.

[6] M. Visser, N. J. Van Eck, and L. Waltman, "Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, crossref, and microsoft academic," *Quantitative science studies*, vol. 2, no. 1, pp. 20–41, 2021.

[7] National Institutes of Health, *2023 NIH data management and sharing policy*, `https://oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/2023-nih-data-management-sharing-policy`, Accessed: 2023-3-23, Jan. 2023.

[8] National Science Foundation, *Open data at NSF*, `https://www.nsf.gov/data/`, Accessed: 2023-3-23.

[9] G. Van Westrienen and C. A. Lynch, "Academic institutional repositories," *D-lib Magazine*, vol. 11, no. 9, pp. 1082–9873, 2005.

[10] P. Jain, "New trends and future applications/directions of institutional repositories in academic institutions," *Library Review*, vol. 60, no. 2, pp. 125–141, 2011.

[11] M. C. Levenstein and J. A. Lyle, "Data: Sharing is caring," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 1, pp. 95–103, Mar. 2018.

[12] E. Moss, *Sweeten the search: Discover data for reuse with a tool that links publications to the underlying data*, 2022.

[13] E. Moss and J. Lyle, "Opaque data citation: Actual citation practice and its implication for tracking data use," 2018.

[14] Y. Kim and M. Adler, "Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories," *International journal of information management*, vol. 35, no. 4, pp. 408–418, 2015.

[15] W. D. Chawinga and S. Zinn, "Global perspectives of research data sharing: A systematic literature review," *Library & Information Science Research*, vol. 41, no. 2, pp. 109–122, 2019.

[16] C. L. Borgman, A. Scharnhorst, and M. S. Golshan, "Digital data archives as knowledge infrastructures: Mediating data sharing and reuse," *Journal of the Association for Information Science and Technology*, vol. 70, no. 8, pp. 888–904, 2019.

[17] L. Fan, S. Lafia, L. Li, F. Yang, and L. Hemphill, "DataChat: Prototyping a conversational agent for dataset search and visualization," May 2023. arXiv: `2305.18358 [cs.IR]`.

[18] C. Tenopir, N. M. Rice, S. Allard, *et al.*, "Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide," *PloS one*, vol. 15, no. 3, e0229003, 2020.

[19] M. Boeckhout, G. A. Zielhuis, and A. L. Bredenoord, "The FAIR guiding principles for data stewardship: Fair enough?" en, *Eur. J. Hum. Genet.*, vol. 26, no. 7, pp. 931–936, Jul. 2018.

[20] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, *et al.*, "The FAIR guiding principles for scientific data management and stewardship," en, *Sci Data*, vol. 3, p. 160 018, Mar. 2016.

[21] B. W. Bishop, C. Hank, J. Webster, and R. Howard, "Scientists' data discovery and reuse behavior:(meta) data fitness for use and the fair data principles," *Proceedings of the association for information science and technology*, vol. 56, no. 1, pp. 21–31, 2019.

[22] I. M. Faniel, R. D. Frank, and E. Yakel, "Context from the data reuser's point of view," *Journal of Documentation*, vol. 75, no. 6, pp. 1274–1297, 2019.

[23] L. Koesten, E. Simperl, T. Blount, E. Kacprzak, and J. Tennison, "Everything you always wanted to know about a dataset: Studies in data summarisation," *International journal of human-computer studies*, vol. 135, p. 102 367, 2020.

[24] J. York, "Seeking equilibrium in data reuse: A study of knowledge satisficing," Ph.D. dissertation, 2022.

[25] A. Dunning, M. De Smaele, and J. Böhmer, "Are the FAIR data principles fair?" *Int. J. Digit. Curation*, vol. 12, no. 2, pp. 177–195, Jan. 2017.

[26] E. Alharbi, R. Skeva, N. Juty, C. Jay, and C. Goble, "Exploring the current practices, costs and benefits of FAIR implementation in pharmaceutical research and development: A qualitative interview study," *Data Intelligence*, vol. 3, no. 4, pp. 507–527, 2021.

[27] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, and E. Motta, "AIDA: A knowledge graph about research dynamics in academia and industry," en, *Quant. Sci. Stud.*, vol. 2, no. 4, pp. 1356–1398, Dec. 2021.

[28] A. Oelen, M. Stocker, and S. Auer, "Creating a scholarly knowledge graph from survey article tables," Dec. 2020. arXiv: `2012.00456 [cs.DL]`.

[29] D. Buscaldi, D. Dess‘i, E. Motta, F. Osborne, and D. R. Recupero, “Mining scholarly data for Fine-Grained knowledge graph construction,” in *DL4KG@ESWC.*, 2019.

[30] H. Turki, M. A. Hadj Taieb, M. Ben Aouicha, G. Fraumann, C. Hauschke, and L. Heller, “Enhancing knowledge graph extraction and validation from scholarly publications using bibliographic metadata,” en, *Front Res Metr Anal*, vol. 6, p. 694 307, May 2021.

[31] Z. W. Nie, Y. J. Liu, L. Y. Yang, S. N. Li, and F Pan, “Construction and application of materials knowledge graph based on author disambiguation: Revisiting the evolution of LiFePO4,” *ADVANCED ENERGY MATERIALS*, vol. 11, no. 16, 2021.

[32] J. C. Wu, Y Jiang, X Chen, L. Y. Guo, X. T. Wei, and X. Y. Yang, ““the canton canon” digital library based on knowledge graph - taking the revolutionary archives of canton in the republic of china as an example,” in *2021 10TH INTERNATIONAL CONFERENCE ON EDUCATIONAL AND INFORMATION TECHNOLOGY (ICEIT 2021)*, 2021, pp. 171–179.

[33] M. Stocker, P. Paasonen, M. Fiebig, M. A. Zaidan, and A. Hardisty, “Curating scientific information in knowledge infrastructures,” en, *Data Sci. J.*, vol. 17, Sep. 2018.

[34] V. G. Meister, “Towards a knowledge graph for a research group with focus on qualitative analysis of scholarly papers,” in *SemSci@ISWC.*, 2017.

[35] A. Oelen, M. Stocker, and S. Auer, “Crowdsourcing scholarly discourse annotations,” in *26th International Conference on Intelligent User Interfaces*, ser. IUI ’21, College Station, TX, USA: Association for Computing Machinery, Apr. 2021, pp. 464–474.

[36] K. L. Hanson, T. DiLauro, and M. Donoghue, “The rmap project: Capturing and preserving associations amongst multi-part distributed publications,” in *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL ’15, Knoxville, Tennessee, USA: Association for Computing Machinery, 2015, pp. 281–282, isbn: 9781450335942. doi: `10.1145/2756406.2756952`. [Online]. Available: `https://doi-org.proxy.lib.umich.edu/10.1145/2756406.2756952`.

[37] O. Karras, E. C. Groen, J. A. Khan, and S. Auer, “Researcher or crowd member? why not both! the open research knowledge graph for applying and communicating CrowdRE research,” Aug. 2021. arXiv: `2108.05085 [cs.DL]`.

[38] E. I. Djebbar and G. Belalem, “Tasks scheduling and resource allocation for high data management in scientific cloud computing environment,” in *Mobile, Secure, and Programmable Networking*, Springer International Publishing, 2016, pp. 16–27.

[39] Neo4j, Inc., *NEO4J GRAPH DATA PLATFORM | Blazing-Fast graph, petabyte scale*, `https://neo4j.com/`, Accessed: 2023-4-10.

[40] D. Su, Y. Xu, G. I. Winata, *et al.*, “Generalizing question answering system with pre-trained language model fine-tuning,” in *Proceedings of the 2nd workshop on machine reading for question answering*, 2019, pp. 203–211.

[41] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," *arXiv [cs.CL]*, 2019. [Online]. Available: `http://arxiv.org/abs/1909.08053`.

[42] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A bibliometric review of large language models research from 2017 to 2023," Apr. 2023. arXiv: `2304.02020 [cs.DL]`.

[43] OpenAI, "GPT-4 technical report," Mar. 2023. arXiv: `2303.08774 [cs.CL]`.

[44] Y. Shen, L. Heacock, J. Elias, *et al.*, "ChatGPT and other large language models are double-edged swords," en, *Radiology*, p. 230 163, Jan. 2023.

[45] W. X. Zhao, K. Zhou, J. Li, *et al.*, *A survey of large language models*, 2023. arXiv: `2303.18223 [cs.CL]`. [Online]. Available: `http://arxiv.org/abs/2303.18223`.

[46] M.-Y. Day and S.-R. Shaw, "AI customer service system with pre-trained language and response ranking models for university admissions," in *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, Aug. 2021, pp. 395–401.

[47] R. Harmouche, A. Lochbihler, F. Thibault, G. De Luca, C. Proulx, and J. L. Hovdebo, "A virtual assistant for cybersickness care," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, Jul. 2020, pp. 384–387.

[48] S. Yu, Y. Chen, and H. Zaidi, "AVA: A financial service chatbot based on deep bidirectional transformers," *Frontiers in Applied Mathematics and Statistics*, vol. 7, 2021.

[49] A. M. Pienta, D. Akmon, J. Noble, L. Hoelter, and S. Jekielek, "A Data-Driven approach to appraisal and selection at a domain data repository," en, *International Journal of Digital Curation*, vol. 12, no. 2, Jun. 2018.

[50] S. Auer and A. Kasprzik, *Towards a Knowledge Graph for Science*, en. Gottfried Wilhelm Leibniz Universität Hannover, 2018.

[51] S Verma, R Bhatia, S Harit, and S Batish, "Scholarly knowledge graphs through structuring scholarly communication: A review," *COMPLEX & INTELLIGENT SYSTEMS*, 2023.

[52] S. M. Swanberg, "Inter-university consortium for political and social research (icpsr)," *Journal of the Medical Library Association: JMLA*, vol. 105, no. 1, p. 106, 2017.

[53] D. Adkins, D. Akmon, J. T. Alexander, G. Alter, and G. C. Alter, "Inter-university consortium for political and social research (icpsr)," 1790.

[54] H. Eulau, *Crossroads of social science: The ICPSR 25th anniversary volume*. Algora Publishing, 2007.

[55] D. Shahi and D. Shahi, "Apache solr: An introduction," *Apache Solr: A practical approach to enterprise search*, pp. 1–9, 2015.

[56] S. Lafia, A. Million, and L. Hemphill, "Exploratory and directed search strategies at a social science data archive," *IASSIST Quarterly*, vol. 48, no. 1, 2024.

[57] K. M. Gregory, H. Cousijn, P. Groth, A. Scharnhorst, and S. Wyatt, “Understanding data search as a socio-technical practice,” *Journal of Information Science*, vol. 46, no. 4, pp. 459–475, 2020.

[58] K. Gregory, P. Groth, H. Cousijn, A. Scharnhorst, and S. Wyatt, “Searching data: A review of observational data retrieval practices in selected disciplines,” *Journal of the Association for Information Science and Technology*, vol. 70, no. 5, pp. 419–432, 2019.

[59] Q. Li, C. Liu, J. Hou, and P. Wang, “Affective memories and perceived value: Motivators and inhibitors of the data search-access process,” *Journal of Documentation*, vol. 79, no. 5, pp. 1236–1264, 2023.

[60] K. Gregory, P. Groth, A. Scharnhorst, and S. Wyatt, “Lost or found? discovering data needed for research,” en, *Harvard Data Science Review*, 2020.

[61] A. Pienta, D. Akmon, J. Noble, L. Hoelter, and S. Jekielek, “A data-driven approach to appraisal and selection at a domain data repository,” *International journal of digital curation*, vol. 12, no. 2, p. 362, 2017.

[62] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, “Microsoft academic graph: When experts are not enough,” en, *Quant. Sci. Stud.*, vol. 1, no. 1, pp. 396–413, Feb. 2020.

[63] OpenAlex, *Topics assigned to works*, Accessed: 2024-04-21, 2024. [Online]. Available: `https://docs.openalex.org/api-entities/topics`.

[64] C. Figueroa, I. Vagliano, O. R. Rocha, and M. Morisio, “A systematic literature review of linked data-based recommender systems,” *Concurrency and Computation: Practice and Experience*, vol. 27, no. 17, pp. 4659–4684, 2015.

[65] Y. Chen, W. Wang, and Z. Liu, “Keyword-based search and exploration on databases,” in *2011 IEEE 27th International Conference on Data Engineering*, IEEE, 2011, pp. 1380–1383.

[66] D. Tümer, M. A. Shah, and Y. Bitirim, “An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia,” in *2009 Fourth International Conference on Internet Monitoring and Protection*, IEEE, 2009, pp. 51–55.

[67] J. Xian, T. Teofili, R. Pradeep, and J. Lin, “Vector search with openai embeddings: Lucene is all you need,” in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 1090–1093.

[68] H. Zhang, Y. Wang, Q. Chen, *et al.*, “Model-enhanced vector index,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[69] J. J. Pan, J. Wang, and G. Li, “Survey of vector database management systems,” *arXiv preprint arXiv:2310.14021*, 2023.

[70] J. Vertesi and P. Dourish, “The value of data: Considering the context of production in data economies,” in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, ser. CSCW ’11, Hangzhou, China: Association for Computing Machinery, Mar. 2011, pp. 533–542.

[71] G. Silvello, “Theory and practice of data citation,” *Journal of the Association for Information Science and Technology*, vol. 69, no. 1, pp. 6–20, 2018.

[72] S. R. Carroll, E. Herczog, M. Hudson, K. Russell, and S. Stall, *Operationalizing the CARE and FAIR principles for indigenous data futures*, 2021.

[73] A. Hasnain and D. Rebholz-Schuhmann, "Assessing FAIR data principles against the 5-star open data principles," in *The Semantic Web: ESWC 2018 Satellite Events*, Springer International Publishing, 2018, pp. 469–477.

[74] P. Manghi, A. Mannocci, F. Osborne, D. Sacharidis, A. Salatino, and T. Vergoulis, "New trends in scientific knowledge graphs and research impact assessment," en, *Quant. Sci. Stud.*, vol. 2, no. 4, pp. 1296–1300, Dec. 2021.

[75] S. Corrall, M. A. Kennan, and W. Afzal, "Bibliometrics and research data management services: Emerging trends in library support for research," *Libr. Trends*, vol. 61, no. 3, pp. 636–674, 2013.

[76] C. L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*, en. MIT Press, Feb. 2017.

[77] W. K. Michener, "Ten simple rules for creating a good data management plan," en, *PLoS Comput. Biol.*, vol. 11, no. 10, e1004525, Oct. 2015.

[78] C. Tenopir, S. Allard, K. Douglass, *et al.*, "Data sharing by scientists: Practices and perceptions," en, *PLoS One*, vol. 6, no. 6, e21101, Jun. 2011.

[79] L. Corti, V. Van den Eynden, L. Bishop, and M. Woollard, *Managing and sharing research data: A guide to good practice*. Sage, 2019.

[80] M. H. Cragin, C. L. Palmer, J. R. Carlson, and M. Witt, "Data sharing, small science and institutional repositories," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 368, no. 1926, pp. 4023–4038, Sep. 2010.

[81] C. A. Lee, *I, Digital: Personal Collections in the Digital Era*, en. Society of American Archivists, 2011.

[82] J. M. Ray, *Research Data Management: Practical Strategies for Information Professionals*, en. Purdue University Press, Nov. 2013.

[83] A. H. Renear, S. Sacchi, and K. M. Wickett, "Definitions of dataset in the scientific and technical literature," en, *Proc. Am. Soc. Inf. Sci. Technol.*, vol. 47, no. 1, pp. 1–4, Nov. 2010.

[84] A. Treloar, "The research data alliance: Globally co-ordinated action against barriers to data publishing and sharing," en, *Learn. Publ.*, vol. 27, no. 5, pp. 9–13, Sep. 2014.

[85] H. Pampel, P. Vierkant, F. Scholze, *et al.*, "Making research data repositories visible: The re3data.org registry," en, *PLoS One*, vol. 8, no. 11, e78080, Nov. 2013.

[86] M. Duke and A. Ball, "How to cite datasets and link to publications: A report of the digital curation centre," en, in *23rd International CODATA Conference*, Oct. 2012.

[87] P. Daybell, A. Payant, E. Woolcott, B. Skeen, A.-M. Arnljots, and K. Meyer, "A tale of two levels: Analyzing the discoverability and impact of Item-Level description in EAD finding aids," *Journal of Archival Organization*, vol. 18, no. 3-4, pp. 101–129, Oct. 2021.

[88] S. Higgins, C. Hilton, and L. Dafis, "Archives context and discovery: Rethinking arrangement and description for the digital age," in *2nd Annual Conference of the International Council on Archives: Archives and Cultural Industries*, 2014.

[89] H. Mooney, "Citing data sources in the social sciences: Do authors do it?" en, *Learn. Publ.*, vol. 24, no. 2, pp. 99–108, Apr. 2011.

[90] D. Brickley, M. Burgess, and N. Noy, "Google dataset search: Building a search engine for datasets in an open web ecosystem," en, in *The World Wide Web Conference on - WWW '19*, San Francisco, CA, USA: ACM Press, 2019, pp. 1365–1375.

[91] M. Vardigan, P. Heus, and W. Thomas, "Data documentation initiative: Toward a standard for the social sciences," en, *International Journal of Digital Curation*, vol. 3, no. 1, pp. 107–113, Dec. 2008.

[92] J. Brase, "DataCite - a global registration agency for research data," in *2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, Nov. 2009, pp. 257–261.

[93] J. C. Wallis, E. Rolando, and C. L. Borgman, "If we share data, will anyone use them? data sharing and reuse in the long tail of science and technology," en, *PLoS One*, vol. 8, no. 7, e67332, Jul. 2013.

[94] A. Yoon, "End users' trust in data repositories: Definition and influences on trust development," en, *Arch. Sci.*, vol. 14, no. 1, pp. 17–34, Mar. 2014.

[95] E. T. Borer, E. W. Seabloom, M. B. Jones, and M. Schildhauer, "Some simple guidelines for effective data management," *Bulletin of the Ecological Society of America*, vol. 90, no. 2, pp. 205–214, 2009.

[96] D. Hassell, J. Gregory, J. Blower, B. N. Lawrence, and K. E. Taylor, "A data model of the climate and forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1)," en, *Geosci. Model Dev.*, vol. 10, no. 12, pp. 4619–4646, Dec. 2017.

[97] J. Wieczorek, D. Bloom, R. Guralnick, *et al.*, "Darwin core: An evolving community-developed biodiversity data standard," en, *PLoS One*, vol. 7, no. 1, e29715, Jan. 2012.

[98] C. M. Chen and C Chang, "A chinese ancient book digital humanities research platform to support digital humanities research," *ELECTRONIC LIBRARY*, vol. 37, no. 2, pp. 314–336, 2019.

[99] C. Guedes, B. Giesteira, and S. Nunes, "Designing user interaction with linked data in historical archives," *J. Comput. Cult. Herit.*, vol. 15, no. 3, Sep. 2022.

[100] K. M. Fear, "Measuring and anticipating the impact of data reuse," Ph.D. dissertation, University of Michigan, 2013.

[101] L. M. Federer, "Who, what, when, where, and why? quantifying and understanding biomedical data reuse," Ph.D. dissertation, University of Maryland, 2019.

[102] I. V. Pasquetto, C. L. Borgman, and M. F. Wofford, "Uses and reuses of scientific data: The data creators' advantage," en, *Harvard Data Science Review*, vol. 1, no. 2, Nov. 2019.

[103] D. Pride and P. Knoth, “An authoritative approach to citation classification,” in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, ser. JCDL ’20, Virtual Event, China: Association for Computing Machinery, Aug. 2020, pp. 337–340.

[104] C. L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*, en. Cambridge, MA: MIT Press, Feb. 2015.

[105] T. Krämer, A. Papenmeier, Z. Carevic, D. Kern, and B. Mathiak, “Data-seeking behaviour in the social sciences,” *International Journal on Digital Libraries*, vol. 22, pp. 175–195, 2021.

[106] M. F. Wofford, B. M. Boscoe, C. L. Borgman, I. V. Pasquetto, and M. S. Golshan, “Jupyter notebooks as discovery mechanisms for open science: Citation practices in the astronomy community,” *Comput. Sci. Eng.*, vol. 22, no. 1, pp. 5–15, Jan. 2020.

[107] J. Brase, M. Lautenschlager, and I. Sens, “The tenth anniversary of assigning DOI names to scientific data and a five year history of DataCite,” *DLib Mag.*, vol. 21, no. 1/2, Jan. 2015.

[108] N. Paskin, “Digital object identifiers for scientific data,” en, *Data Sci. J.*, vol. 4, pp. 12–20, 2005.

[109] M. Altman, C. Borgman, M. Crosas, and M. Matone, “Data citation synthesis group: Joint declaration of data citation principles,” *Bulletin of the Association for Information Science and Technology*, vol. 41, no. 3, pp. 43–45, 2015.

[110] C. Tauchert, P. Buxmann, and J. Lambinus, “Crowdsourcing data science: A qualitative analysis of organizations’ usage of kaggle competitions,” 2020.

[111] M. Z. Al-Taie, N. Salim, and A. I. Obasa, “Successful data science projects: Lessons learned from kaggle competition,” *Kurdistan Journal of Applied Research*, vol. 2, no. 3, pp. 40–49, 2017.

[112] D. A. Benson, M. Cavanaugh, K. Clark, *et al.*, “Genbank,” *Nucleic acids research*, vol. 41, no. D1, pp. D36–D42, 2012.

[113] P. Brooker, “Sesar and nextgen: Investing in new paradigms,” *The Journal of Navigation*, vol. 61, no. 2, pp. 195–208, 2008.

[114] M. Diepenbroek, H. Grobe, M. Reinke, *et al.*, “Pangaea—an information system for environmental sciences,” *Computers & Geosciences*, vol. 28, no. 10, pp. 1201–1210, 2002.

[115] C. Brown, “The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance,” *Journal of the American Society for Information Science and Technology*, vol. 54, no. 10, pp. 926–938, Aug. 2003.

[116] G. King, “Replication, replication,” *PS: Political Science & Politics*, vol. 28, no. 3, pp. 444–452, Sep. 1995.

[117] I. V. Pasquetto, B. M. Randles, and C. L. Borgman, “On the reuse of scientific data,” *Data Science Journal*, vol. 16, no. 8, Mar. 2017.

[118] A. K. Thomer, “Integrative data reuse at scientifically significant sites: Case studies at yellowstone national park and the la brea tar pits,” en, *J. Assoc. Inf. Sci. Technol.*, Feb. 2022.

[119] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: data-intensive scientific discovery*. Redmond, Washington: Microsoft Research, 2009.

[120] C. L. Palmer, N. M. Weber, and M. H. Cragin, "The analytic potential of scientific data: Understanding re-use value," en, *Proceedings of the American Society for Information Science and Technology*, vol. 48, no. 1, pp. 1–10, 2011.

[121] National Academy of Sciences, *Facilitating Interdisciplinary Research*, en. National Academies Press, May 2005.

[122] L. Darden and N. Maull, "Interfield theories," *Philos. Sci.*, vol. 44, no. 1, pp. 43–64, Mar. 1977.

[123] I. Gökalp, "On the dynamics of controversies in a borderland scientific domain: The case of turbulent combustion," *Soc. Sci. Inf.*, vol. 26, no. 3, pp. 551–576, Sep. 1987.

[124] S. L. Star and J. R. Griesemer, *Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39*, 1989.

[125] Data Citation Synthesis Group, "Joint declaration of data citation principles," Tech. Rep., 2014.

[126] M. Fenner, M. Crosas, J. Grethe, *et al.*, "A data citation roadmap for scholarly data repositories," *Scientific Data*, vol. 6, no. 1, Dec. 2016.

[127] H. Cousijn, P. Feeney, D. Lowenberg, E. Presani, and N. Simons, "Bringing citations and usage metrics together to make data count," *Data Science Journal*, vol. 18, no. 1, 2019.

[128] E. A. Leicht, G Clarkson, K Shedden, and M. E. J. Newman, "Large-scale structure of time evolving citation networks," en, *European Physical Journal B*, vol. 59, no. 1, pp. 75–83, 2007.

[129] C Chen, "Science mapping: A systematic review of the literature," *Journal of data and information science*, 2017.

[130] M. V. Tomasello, G. Vaccario, and F. Schweitzer, "Data-driven modeling of collaboration networks: A cross-domain analysis," *EPJ Data Science*, vol. 6, no. 1, p. 22, Sep. 2017.

[131] M. E. J. Newman, "Who is the best connected scientist? a study of scientific coauthorship networks," in *Complex Networks*, E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 337–370.

[132] R. Bose and J. Frew, "Lineage retrieval for scientific data processing: A survey," *ACM Computing Surveys*, vol. 37, no. 1, pp. 1–28, 2005.

[133] P. Buneman, D. Dosso, M. Lissandrini, and G. Silvello, "Data citation and the citation graph," en, *Quantitative Science Studies*, vol. 2, no. 4, pp. 1399–1422, Dec. 2021.

[134] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," en, *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, Jul. 1973.

[135] H. D. White and K. W. McCain, "Visualizing a discipline: An author Co-Citation analysis of information science, 1972–1995," *Journal of the Association for Information Science and Technology*, 1998.

[136] D. J. D. S. Price and D. Beaver, “Collaboration in an invisible college,” *American Psychologist*, vol. 21, no. 11, pp. 1011–1018, 1966.

[137] D. Crane, “Social structure in a group of scientists: A test of the “invisible college” hypothesis,” in *Social Networks*, Elsevier, 1977, pp. 161–178.

[138] G. Franck, “Scientific Communication–A vanity fair?” *Science*, vol. 286, no. 5437, pp. 53–55, 1999.

[139] A. Varga, “Shorter distances between papers over time are due to more cross-field references and increased citation rate to higher-impact papers,” en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 44, pp. 22 094–22 099, Oct. 2019.

[140] A. L. Porter and I. Rafols, “Is science becoming more interdisciplinary? measuring and mapping six research fields over time,” *Scientometrics*, vol. 81, no. 3, p. 719, Apr. 2009.

[141] L. A. Orthia, M. McKinnon, J. N. Viana, and G. Walker, “Reorienting science communication towards communities,” *Journal of Science Communication*, vol. 20, no. 03, A12, May 2021.

[142] P. B. Heidorn, “Shedding light on the dark data in the long tail of science,” *Library Trends*, vol. 57, no. 2, pp. 280–299, 2008.

[143] A. Sands, C. L. Borgman, L. Wynholds, and S. Traweek, “Follow the data: How astronomers use and reuse data,” en, *Proceedings of the American Society for Information Science and Technology*, vol. 49, no. 1, pp. 1–3, 2012.

[144] A. K. Thomer, M. B. Twidale, and M. J. Yoder, “Transforming taxonomic interfaces,” en, *Proc. ACM Hum. Comput. Interact.*, vol. 2, no. CSCW, pp. 1–23, Nov. 2018.

[145] A. S. Zimmerman, “New knowledge from old data: The role of standards in the sharing and reuse of ecological data,” *Science, Technology, & Human Values*, vol. 33, no. 5, pp. 631–652, Sep. 2008.

[146] M. Altman, C. Borgman, M. Crosas, and M. Matone, “An introduction to the joint principles for data citation,” *Bulletin of the Association for Information Science and Technology*, vol. 41, no. 3, pp. 43–45, 2015.

[147] K. Helbig, B. Hausstein, and R. Toepfer, “Supporting data citation: Experiences and best practices of a doi allocation agency for social sciences,” *Journal of Librarianship and Scholarly Communication*, vol. 3, no. 2, 2015.

[148] J. Brase, “Making data citeable: Datacite,” in *Opening Science*, Springer, Cham, 2014, pp. 327–329.

[149] S. Pröll and A. Rauber, “Scalable data citation in dynamic, large databases: Model and reference implementation,” in *2013 IEEE International Conference on Big Data*, 2013, pp. 307–312. doi: `10.1109/BigData.2013.6691588`.

[150] S. Pröll and A. Rauber, “Scalable data citation in dynamic, large databases: Model and reference implementation,” in *2013 IEEE International Conference on Big Data*, IEEE, 2013, pp. 307–312.

[151] D. Dimitrov, E. Baran, and D. Wegener, “Making data citable-a web-based system for the registration of social and economics science data.,” in *WEBIST*, 2013, pp. 155–159.

[152] S. Proell and A. Rauber, “A scalable framework for dynamic data citation of arbitrary structured data.,” in *DATA*, 2014, pp. 223–230.

[153] D. Kern and B. Mathiak, “Are there any differences in data set retrieval compared to well-known literature retrieval?” In *Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, September 14-18, 2015, Proceedings 19*, Poznań, Poland: Springer, 2015, pp. 197–208.

[154] L. A. Wynholds, J. C. Wallis, C. L. Borgman, A. Sands, and S. Traweek, “Data, data use, and scientific inquiry: Two case studies of data practices,” in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, ser. JCDL ’12, Washington, DC, USA: Association for Computing Machinery, Jun. 2012, pp. 19–22.

[155] C. Jiao and P. T. Darch, “The role of the data paper in scholarly communication,” en, *Proc. Assoc. Inf. Sci. Technol.*, vol. 57, no. 1, Oct. 2020.

[156] K. Gregory, P. Groth, H. Cousijn, A. Scharnhorst, and S. Wyatt, “Searching data: A review of observational data retrieval practices in selected disciplines,” en, *J Assoc Inf Sci Technol*, vol. 70, no. 5, pp. 419–432, May 2019.

[157] K. Li and C. Jiao, “The data paper as a sociolinguistic epistemic object: A content analysis on the rhetorical moves used in data paper abstracts,” en, *J. Assoc. Inf. Sci. Technol.*, vol. 73, no. 6, pp. 834–846, Jun. 2022.

[158] M. S. Mayernik, D. L. Hart, K. E. Maull, and N. M. Weber, “Assessing and tracing the outcomes and impact of research infrastructures,” en, *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 6, pp. 1341–1359, Jun. 2017.

[159] N. Robinson-García, E. Jiménez-Contreras, and D. Torres-Salinas, “Analyzing data citation practices using the data citation index,” *Journal of the Association for Information Science and Technology*, vol. 67, no. 12, pp. 2964–2975, 2016.

[160] H. Mooney and M. P. Newton, “The anatomy of a data citation: Discovery, reuse, and credit,” *Journal of Librarianship and Scholarly Communication*, vol. 1, no. 1, 2012.

[161] M. A. Parsons, R. E. Duerr, and M. B. Jones, “The history and future of data citation in practice,” *Data Science Journal*, vol. 18, no. 1, 2019.

[162] A. Alawini, S. B. Davidson, W. Hu, and Y. Wu, “Automating data citation in citedb,” *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 1881–1884, 2017.

[163] S. B. Davidson, D. Deutch, T. Milo, and G. Silvello, “A model for fine-grained data citation.,” in *CIDR*, 2017.

[164] G. Silvello, “Learning to cite framework: How to automatically construct citations for hierarchical data,” *Journal of the Association for Information Science and Technology*, vol. 68, no. 6, pp. 1505–1524, 2017.

[165] S. B. Davidson, P. Buneman, D. Deutch, T. Milo, and G. Silvello, "Data citation: A computational challenge," in *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2017, pp. 1–4.

[166] M. M. Force and N. J. Robinson, "Encouraging data citation and discovery with the data citation index," *Journal of computer-aided molecular design*, vol. 28, no. 10, pp. 1043–1048, 2014.

[167] E. Moss and J. Lyle, "Opaque data citation: Actual citation practice and its implication for tracking data use," Poster presented at the 13th International Digital Curation Conference, 2018. [Online]. Available: `https://deepblue.lib.umich.edu/handle/2027.42/142393`.

[168] F Zhang, Z. Y. Li, D. H. Peng, and J. W. Cheng, "RDF for temporal data management - a survey," *Earth Sci. Inf.*, vol. 14, no. 2, pp. 563–599, 2021.

[169] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson, "Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the european open science cloud," *Inf. Serv. Use*, vol. 37, no. 1, pp. 49–56, 2017.

[170] P. H. P. Jati, Y. Lin, S. Nodehi, D. B. Cahyono, and M. van Reisen, "FAIR versus open data: A comparison of objectives and principles," en, *Data Intell.*, vol. 4, no. 4, pp. 867–881, Oct. 2022.

[171] Anna Bernasconi, Alberto S, Veda Storey, Giancarlo Guizzardi, Luiz Santos, *Ontological representation of FAIR principles: A blueprint for FAIRer data sources*, `https://www.researchgate.net/profile/Giancarlo-Guizzardi-2/publication/369850668_Ontological_representation_of_FAIR_principles_A_blueprint_for_FAIRer_data_sources/links/642f51c84e83cd0e2f95b820/Ontological-representation-of-FAIR-principles-A-blueprint-for-FAIRer-data-sources.pdf`, Accessed: 2023-5-27, 2023.

[172] S. Leonelli, R. Lovell, B. W. Wheeler, L. Fleming, and H. Williams, "From FAIR data to fair data use: Methodological data fairness in health-related social media research," *Big Data & Society*, vol. 8, no. 1, p. 20 539 517 211 010 310, Jan. 2021.

[173] A. Sinha, Z. Shen, Y. Song, *et al.*, "An overview of microsoft academic service (MAS) and applications," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion, Florence, Italy: Association for Computing Machinery, May 2015, pp. 243–246.

[174] L. Waltman and N. J. van Eck, "Field-normalized citation impact indicators and the choice of an appropriate counting method," *J. Informetr.*, vol. 9, no. 4, pp. 872–894, Oct. 2015.

[175] D. Shotton, "Semantic publishing: The coming revolution in scientific journal publishing," *Learn. Publ.*, vol. 22, no. 2, pp. 85–94, 2009.

[176] A. Hogan, E. Blomqvist, M. Cochez, *et al.*, "Knowledge graphs," Mar. 2020. arXiv: `2003.02320 [cs.AI]`.

[177] M. Färber and D. Lamprecht, "The data set knowledge graph: Creating a linked open data source for data sets," *Quantitative Science Studies*, vol. 2, no. 4, pp. 1324–1355, 2021.

[178] F. Bensmann, L. Heling, S. Jünger, *et al.*, "An infrastructure for spatial linking of survey data," en, *Data Sci. J.*, vol. 19, Jul. 2020.

[179] E. Arbeláez-Cortés, A. R. Acosta-Galvis, C. Nascimiento, D. Espitia-Reina, A. González-Alvarado, and C. A. Medina, "Knowledge linked to museum specimen vouchers: Measuring scientific production from a major biological collection in colombia," *Scientometrics*, vol. 112, no. 3, pp. 1323–1341, Sep. 2017.

[180] M. Färber, *The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data*, 2019.

[181] M. Y. Jaradeh, A. Oelen, K. E. Farfar, *et al.*, "Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge," in *Proceedings of the 10th International Conference on Knowledge Capture*, ser. K-CAP '19, Marina Del Rey, CA, USA: Association for Computing Machinery, Sep. 2019, pp. 243–246.

[182] M. Baglioni, P. Manghi, and A. Mannocci, "Context-Driven discoverability of research data," in *Digital Libraries for Open Knowledge*, Springer International Publishing, 2020, pp. 197–211.

[183] I. Cachola, K. Lo, A. Cohan, and D. Weld, "TLDR: Extreme summarization of scientific documents," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, 2020.

[184] T. Hope, A. Amini, D. Wadden, *et al.*, "Extracting a knowledge base of mechanisms from COVID-19 papers," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 4489–4503.

[185] A. Devaraju, M. Mokrane, L. Cepinskas, *et al.*, "From conceptualization to implementation: FAIR assessment of research data objects," en, *Data Sci. J.*, vol. 20, Feb. 2021.

[186] M. Stocker, T. Heger, A. M. Schweidtmann, *et al.*, "SKG4EOSC - scholarly knowledge graphs for EOSC: Establishing a backbone of knowledge graphs for FAIR scholarly information in EOSC," *Riogrande Odontol.*, vol. 8, e83789, Mar. 2022.

[187] R. Chandrakar, "Digital object identifier system: An overview," *The Electronic Library*, vol. 24, no. 4, pp. 445–452, Jan. 2006.

[188] N. Juty, N. Le Novère, and C. Laibe, "Identifiers.org and MIRIAM registry: Community resources to provide persistent identification," en, *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D580–6, Jan. 2012.

[189] P Leach, M Mealling, and R Salz, "A universally unique IDentifier (UUID) URN namespace," Tech. Rep. rfc4122, Jul. 2005.

[190] H. Hussein, A. Oelen, O. Karras, and S. Auer, "KGMM – a maturity model for scholarly knowledge graphs based on intertwined Human-Machine collaboration," Nov. 2022. arXiv: `2211.12223 [cs.DL]`.

[191] R. N. J. Graham, R. W. Perriss, and A. F. Scarsbrook, "DICOM demystified: A review of digital file formats and their use in radiological practice," en, *Clin. Radiol.*, vol. 60, no. 11, pp. 1133–1140, Nov. 2005.

[192] A. Oelen, M. Y. Jaradeh, and K. E. Farfar, "Comparing research contributions in a scholarly knowledge graph," in *SciKnow@K-CAP*, 2019.

[193] S. Hanif and N. F. Waraich, "Link your metadata: Integrating library collection through open refine," *Pakistan Library & Information Science Journal*, vol. 53, no. 2, 2022.

[194] M. Färber and L. Ao, "The microsoft academic knowledge graph enhanced: Author name disambiguation, publication classification, and embeddings," en, *Quant. Sci. Stud.*, pp. 1–48, Mar. 2022.

[195] R. Wang, Y. Yan, J. Wang, *et al.*, "AceKG: A large-scale knowledge graph for academic data mining," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ser. CIKM '18, Torino, Italy: Association for Computing Machinery, Oct. 2018, pp. 1487–1490.

[196] F. H. Shang, Q. Y. Ding, R. S. Du, M. J. Cao, and H. Y. Chen, "Construction and application of the user behavior knowledge graph in software platforms," *JOURNAL OF WEB ENGINEERING*, vol. 20, no. 2, pp. 387–411, 2021.

[197] H. N. Chen and X. W. Luo, "An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing," *ADVANCED ENGINEERING INFORMATICS*, vol. 42, 2019.

[198] J. Liu, J. Ren, W. Zheng, L. Chi, I. Lee, and F. Xia, "Web of scholars: A scholar knowledge graph," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20, Virtual Event, China: Association for Computing Machinery, Jul. 2020, pp. 2153–2156.

[199] J. Postel, "Transmission control protocol," Tech. Rep., 1981.

[200] J. Blanke and T. Riechert, "Towards an RDF knowledge graph of scholars from early modern history," Sep. 2020. arXiv: `2009.06337 [cs.DL]`.

[201] D Dessi, F Osborne, D. R. Recupero, D Buscaldi, and E Motta, "Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain," *FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE*, vol. 116, pp. 253–264, 2021.

[202] S. Vahdati, G. Palma, R. J. Nath, C. Lange, S. Auer, and M.-E. Vidal, "Unveiling scholarly communities over knowledge graphs," Jul. 2018. arXiv: `1807.06816 [cs.DL]`.

[203] L Penev, M Dimitrova, V Senderov, *et al.*, "OpenBiodiv: A knowledge graph for Literature-Extracted linked open data in biodiversity science," *PUBLICATIONS*, vol. 7, no. 2, 2019.

[204] G. Heidari, A. Ramadan, M. Stocker, and S. Auer, "Demonstration of faceted search on scholarly knowledge graphs," Jul. 2021. arXiv: `2107.05738 [cs.DL]`.

[205] J. J. Wu, Z. Q. Wei, D. N. Jia, X Dou, H Tang, and N. N. Li, "Constructing marine expert management knowledge graph based on Trellisnet-CRF," *PEERJ COMPUTER SCIENCE*, vol. 8, 2022.

[206] L. Vogt, J. D'Souza, M. Stocker, and S. Auer, "Toward representing research contributions in scholarly knowledge graphs using knowledge graph cells," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, ser. JCDL '20, Virtual Event, China: Association for Computing Machinery, Aug. 2020, pp. 107–116.

[207] A. Brazma, P. Hingamp, J. Quackenbush, *et al.*, "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data," en, *Nat. Genet.*, vol. 29, no. 4, pp. 365–371, Dec. 2001.

[208] C. F. Taylor, N. W. Paton, K. S. Lilley, *et al.*, "The minimum information about a proteomics experiment (MIAPE)," en, *Nat. Biotechnol.*, vol. 25, no. 8, pp. 887–893, Aug. 2007.

[209] B. Haslhofer, A. Isaac, and R. Simon, "Knowledge graphs in the libraries and digital humanities domain," Mar. 2018. arXiv: `1803.03198 [cs.DL]`.

[210] A. Oelen, M. Stocker, and S. Auer, "TinyGenius: Intertwining natural language processing with micro-task crowdsourcing for scholarly knowledge graph creation," in *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, ser. JCDL '22, Cologne, Germany: Association for Computing Machinery, Jun. 2022, pp. 1–5.

[211] A Oelen, M. Y. Jaradeh, M Stocker, and S Auer, "Generate FAIR literature surveys with scholarly knowledge graphs," Jun. 2020. arXiv: `2006.01747 [cs.DL]`.

[212] L. Koesten, K. Gregory, P. Groth, and E. Simperl, "Talking datasets – understanding data sensemaking behaviours," en, *Int. J. Hum. Comput. Stud.*, vol. 146, no. 102562, p. 102 562, Feb. 2021.

[213] J. York, *Seeking equilibrium in data reuse: A study of knowledge satisficing*, en, 2022.

[214] D. W. Hook, S. J. Porter, and C. Herzog, "Dimensions: Building context for search and evaluation," *Frontiers in Research Metrics and Analytics*, vol. 3, p. 23, 2018. doi: `https://doi.org/10.3389/frma.2018.00023`.

[215] C. D. Corley and R. Mihalcea, "Measuring the semantic similarity of texts," in *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, 2005, pp. 13–18.

[216] X. Huang, "Developing a cross-disciplinary typology of topical relevance relationships as the basis for a topic-oriented information architecture," *Advances in Classification Research Online*, vol. 20, no. 1, 2009.

[217] B. Boyce, "Beyond topicality: A two stage view of relevance and the retrieval process," *Information Processing & Management*, vol. 18, no. 3, pp. 105–109, 1982.

[218] A. G. Maguitman, R. L. Cecchini, C. M. Lorenzetti, and F. Menczer, "Using topic ontologies and semantic similarity data to evaluate topical search," in *XXXVI Conferencia Latinoamericana de Informática. Centro Latinoamericano de Estudios en Informática, Facultad Politécnica–Universidad Nacional de Asunción and Universidad Autónoma de Asunción, Asunción, Paraguay*, 2010.

[219] M. J. Welch, U. Schonfeld, D. He, and J. Cho, “Topical semantics of twitter links,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 327–336.

[220] A. Das and A. Kannan, “Discovering topical aspects in microblogs,” in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 860–871.

[221] M. Rowe, C. Wagner, M. Strohmaier, and H. Alani, “Measuring the topical specificity of online communities,” in *Extended Semantic Web Conference*, Springer, 2013, pp. 472–486.

[222] T. Highfield, “Talking of many things: Using topical networks to study discussions in social media,” *Journal of technology in human services*, vol. 30, no. 3-4, pp. 204–218, 2012.

[223] X. Wang, A. McCallum, and X. Wei, “Topical n-grams: Phrase and topic discovery, with an application to information retrieval,” in *Seventh IEEE international conference on data mining (ICDM 2007)*, IEEE, 2007, pp. 697–702.

[224] J. Bischof and E. M. Airoldi, “Summarizing topical content with word frequency and exclusivity,” in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012, pp. 201–208.

[225] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, “Topical word embeddings,” in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[226] “About icpsr.” (n.d.), [Online]. Available: `https://www.icpsr.umich.edu/web/pages/about/` (visited on 11/02/2021).

[227] “Subject thesaurus.” (n.d.), [Online]. Available: `https://www.icpsr.umich.edu/web/ICPSR/thesaurus/index` (visited on 11/02/2021).

[228] H. S. Farber and M. J. White, *Comparison of formal and informal dispute resolution in medical malpractice*, 1996. doi: `10.3886/ICPSR01059.v1`. [Online]. Available: `https://doi.org/10.3886/ICPSR01059.v1`.

[229] C. Fellbaum, “Wordnet,” in *Theory and applications of ontology: computer applications*, Springer, 2010, pp. 231–243.

[230] G. A. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.

[231] “What is wordnet?” (n.d.), [Online]. Available: `https://wordnet.princeton.edu/` (visited on 12/08/2021).

[232] C. Leacock and M. Chodorow, “Combining local context and wordnet similarity for word sense identification,” *WordNet: An electronic lexical database*, vol. 49, no. 2, pp. 265–283, 1998.

[233] Z. Wu and M. Palmer, “Verb semantics and lexical selection,” *arXiv preprint cmp-lg/9406033*, 1994.

[234] P. Resnik, “Using information content to evaluate semantic similarity in a taxonomy,” *arXiv preprint cmp-lg/9511007*, 1995.

[235] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.

[236] T. Pedersen, S. Patwardhan, J. Michelizzi, *et al.*, "Wordnet:: Similarity-measuring the relatedness of concepts.," in *AAAI*, vol. 4, 2004, pp. 25–29.

[237] R. Navigli and F. Martelli, "An overview of word and sense similarity," *Natural Language Engineering*, vol. 25, no. 6, pp. 693–714, 2019. doi: `https://doi.org/10.1017/S1351324919000305`.

[238] K. Abdalgader and A. Skabar, "Short-text similarity measurement using word sense disambiguation and synonym expansion," in *Australasian joint conference on artificial intelligence*, Springer, 2010, pp. 435–444. doi: `https://doi.org/10.1007/978-3-642-17432-2_44`.

[239] C. Ho, M. A. A. Murad, R. A. Kadir, and S. C. Doraisamy, "Word sense disambiguation-based sentence similarity," in *Coling 2010: Posters*, 2010, pp. 418–426. [Online]. Available: `https://aclanthology.org/C10-2048.pdf`.

[240] E. W. Dijkstra *et al.*, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[241] "Wordnet::similarity::path." (n.d.), [Online]. Available: `https://metacpan.org/pod/WordNet::Similarity::path` (visited on 12/08/2021).

[242] "Source code for nltk.corpus.reader.wordnet." (n.d.), [Online]. Available: `https://www.nltk.org/_modules/nltk/corpus/reader/wordnet.html` (visited on 12/08/2021).

[243] D. Lin *et al.*, "An information-theoretic definition of similarity.," in *Icml*, vol. 98, 1998, pp. 296–304.

[244] "Documentation." (n.d.), [Online]. Available: `https://www.nltk.org/howto/wordnet.html#similarity` (visited on 11/02/2021).

[245] "Natural language toolkit: Wordnet." (n.d.), [Online]. Available: `https://github.com/nltk/nltk/blob/develop/nltk/corpus/reader/wordnet.py` (visited on 12/08/2021).

[246] J. Becsey, L. Berke, and J. R. Callan, "Nonlinear least squares methods: A direct grid search approach," *Journal of Chemical Education*, vol. 45, no. 11, p. 728, 1968.

[247] "Sklearn.metrics.balanced_accuracy_score." (n.d.), [Online]. Available: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html` (visited on 11/02/2021).

[248] H. Wechsler, *Harvard school of public health college alcohol study, 2001*, 2008. doi: `10.3886/ICPSR04291.v2`. [Online]. Available: `https://doi.org/10.3886/ICPSR04291.v2`.

[249] ICPSR, *Data-Related publications*, `https://www.icpsr.umich.edu/web/pages/ICPSR/citations/`, 2023.

[250] Snowflake Inc., *A faster way to build and share data apps*, `https://streamlit.io/`, Accessed: 2023-4-10.

[251] OpenAI, *OpenAI API*, https://openai.com/blog/openai-api, Accessed: 2023-4-10, Sep. 2020.

[252] OpenAI, *GPT-3.5*, https://platform.openai.com/docs/models/gpt-3-5, Accessed: 2023-4-10.

[253] C. Klose, *Streamlit-agraph*, Jan. 2023.

[254] A. Papenmeier, T. Krämer, T. Friedrich, D. Hienert, and D. Kern, "Genuine information needs of social scientists looking for data," *Proceedings of the Association for Information Science and Technology*, vol. 58, no. 1, pp. 292–302, Oct. 2021.

[255] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, en. SAGE Publications, May 2018.

[256] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "GPTs are GPTs: An early look at the labor market impact potential of large language models," Mar. 2023. arXiv: 2303.10130 [econ.GN].

[257] L. Hemphill, A. Thomer, S. Lafia, L. Fan, D. Bleckley, and E. Moss, "A dataset for measuring the impact of research data and their curation," *Scientific Data*, vol. 11, no. 1, p. 442, 2024.

[258] J. Monteiro, F. Sá, and J. Bernardino, "Experimental evaluation of graph databases: Janusgraph, nebula graph, neo4j, and tigergraph," *Applied Sciences*, vol. 13, no. 9, p. 5770, 2023.

[259] D. H. Goh, S.-J. Chen, and S. Tuarob, *Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration: 25th International Conference on Asia-Pacific Digital Libraries, ICADL 2023, Taipei, Taiwan, December 4–7, 2023, Proceedings, Part II*. Springer Nature, 2023, vol. 14458.

[260] C. S. Khoo, E. A. Tan, S.-G. Ng, C.-F. Chan, M. Stanley-Baker, and W.-N. Cheng, "Knowledge graph visualization interface for digital heritage collections: Design issues and recommendations," *Information Technology and Libraries*, vol. 43, no. 1, 2024.

[261] S. Lafia, L. Fan, A. Thomer, and L. Hemphill, "Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network," *Quantitative Science Studies*, pp. 1–21, Oct. 2022, issn: 2641-3337. doi: 10.1162/qss_a_00209. eprint: https://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00209/2055273/qss_a_00209.pdf. [Online]. Available: https://doi.org/10.1162/qss_a_00209.

[262] H. A. Piwowar and T. J. Vision, "Data reuse and the open data citation advantage," *PeerJ*, vol. 1, e175, 2013.

[263] H. Cousijn, T. Habermann, E. Krznarich, and A. Meadows, "Beyond data: Sharing related research outputs to make data reusable.," *Learned Publishing*, vol. 35, no. 1, 2022.

[264] J. Priem, H. Piwowar, and R. Orr, "Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," *arXiv preprint arXiv:2205.01833*, 2022.

[265] L. Fan, W. Hua, L. Li, H. Ling, Y. Zhang, and L. Hemphill, "Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes," *arXiv preprint arXiv:2312.14890*, 2023.

[266] D. Hendrycks, C. Burns, S. Basart, *et al.*, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[267] X. Zhang, C. Li, Y. Zong, Z. Ying, L. He, and X. Qiu, "Evaluating the performance of large language models on gaokao benchmark," *arXiv preprint arXiv:2305.12474*, 2023.

[268] M. Suzgun, N. Scales, N. Schärli, *et al.*, "Challenging big-bench tasks and whether chain-of-thought can solve them," *arXiv preprint arXiv:2210.09261*, 2022.

[269] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, "Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs," *arXiv preprint arXiv:1903.00161*, 2019.

[270] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" *arXiv preprint arXiv:1905.07830*, 2019.

[271] S. Frieder, L. Pinchetti, R.-R. Griffiths, *et al.*, "Mathematical capabilities of chatgpt," *arXiv preprint arXiv:2301.13867*, 2023.

[272] D. S. Johnson, "A catalog of complexity classes," in *Algorithms and complexity*, Elsevier, 1990, pp. 67–161.

[273] J. Achiam, S. Adler, S. Agarwal, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[274] Anthropic, *Claude 2*, 2023. [Online]. Available: `https://www.anthropic.com/news/claude-2`.

[275] OpenAI, *Gpt-3.5 turbo*, 2024. [Online]. Available: `https://platform.openai.com/docs/models/gpt-3-5-turbo`.

[276] Anthropic, *Releasing claude instant 1.2*, 2023. [Online]. Available: `https://www.anthropic.com/news/releasing-claude-instant-1-2`.

[277] Google, *Palm 2*, 2023. [Online]. Available: `https://ai.google/discover/palm2/`.

[278] 01-AI, *Building the next generation of open-source and bilingual llms*, 2023. [Online]. Available: `https://huggingface.co/01-ai/Yi-34B`.

[279] J. Bai, S. Bai, Y. Chu, *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[280] A. Q. Jiang, A. Sablayrolles, A. Mensch, *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[281] M. Javaheripi and S. Bubeck, *Phi-2*, 2023. [Online]. Available: `https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/`.

[282] MosaicML, *Mpt-30b: Raising the bar for open-source foundation models*, 2023. [Online]. Available: `https://www.mosaicml.com/blog/mpt-30b`.

[283] LMSYS, *Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality*, 2023. [Online]. Available: `https://lmsys.org/blog/2023-03-30-vicuna/`.

[284] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks are all you need ii: Phi-1.5 technical report," *arXiv preprint arXiv:2309.05463*, 2023.

[285] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020. [Online]. Available: `http://arxiv.org/abs/2005.14165`.

[286] A. Chowdhery, S. Narang, J. Devlin, *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[287] H. W. Chung, L. Hou, S. Longpre, *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[288] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2022.

[289] W. Hua, L. Fan, L. Li, *et al.*, "War and peace (waragent): Large language model-based multi-agent simulation of world wars," *arXiv preprint arXiv:2311.17227*, 2023.

[290] L. Fan, S. Lafia, L. Li, F. Yang, and L. Hemphill, "Datachat: Prototyping a conversational agent for dataset search and visualization," *arXiv preprint arXiv:2305.18358*, 2023.

[291] Z. Gao, L. Li, S. Ma, Q. Wang, L. Hemphill, and R. Xu, "Examining the potential of chatgpt on biomedical information retrieval: Fact-checking drug-disease associations," *Annals of Biomedical Engineering*, pp. 1–9, 2023.

[292] L. Li, L. Fan, S. Atreja, and L. Hemphill, ""hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media," *arXiv preprint arXiv:2304.10619*, 2023.

[293] J. Wei, Y. Tay, R. Bommasani, *et al.*, "Emergent abilities of large language models," *Trans. Mach. Learn. Res.*, 2022. [Online]. Available: `http://arxiv.org/abs/2206.07682`.

[294] R. Schaeffer, B. Miranda, and S. Koyejo, "Are emergent abilities of large language models a mirage?" *arXiv preprint arXiv:2304.15004*, 2023.

[295] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[296] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[297] B. Wang, X. Deng, and H. Sun, "Iteratively prompt pre-trained language models for chain of thought," *arXiv preprint arXiv:2203.08383*, 2022.

[298] W. Hua and Y. Zhang, "System 1+ system 2= better world: Neural-symbolic chain of logic reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 601–612.

[299] S. Yao, D. Yu, J. Zhao, *et al.*, "Tree of thoughts: Deliberate problem solving with large language models," *arXiv preprint arXiv:2305.10601*, 2023.

[300] M. Besta, N. Blach, A. Kubicek, *et al.*, “Graph of thoughts: Solving elaborate problems with large language models,” *arXiv preprint arXiv:2308.09687*, 2023.

[301] Y. Wang, Z. Jiang, Z. Chen, *et al.*, “Recmind: Large language model powered agent for recommendation,” *arXiv preprint arXiv:2308.14296*, 2023.

[302] G. Kim, P. Baldi, and S. McAleer, “Language models can solve computer tasks,” *arXiv preprint arXiv:2303.17491*, 2023.

[303] Y. Weng, M. Zhu, S. He, K. Liu, and J. Zhao, “Large language models are reasoners with self-verification,” *arXiv preprint arXiv:2212.09561*, 2022.

[304] J. Wei, J. Wei, Y. Tay, *et al.*, “Larger language models do in-context learning differently,” *arXiv preprint arXiv:2303.03846*, 2023.

[305] S. Min, X. Lyu, A. Holtzman, *et al.*, “Rethinking the role of demonstrations: What makes in-context learning work?” *arXiv preprint arXiv:2202.12837*, 2022.

[306] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018. [Online]. Available: `http://arxiv.org/abs/1810.04805`.

[307] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[308] F. Rahutomo, T. Kitasuka, M. Aritsugi, *et al.*, “Semantic cosine similarity,” in *The 7th international student conference on advanced science and technology ICAST*, University of Seoul South Korea, vol. 4, 2012, p. 1.

[309] OpenAI, *Fine-tuning*, Accessed: 2024-04-22, 2024. [Online]. Available: `https://platform.openai.com/docs/guides/fine-tuning`.

[310] R. C. Coile, “Lotka’s frequency distribution of scientific productivity,” *Journal of the american society for information science*, vol. 28, no. 6, pp. 366–370, 1977.

[311] B. Macpherson, “The role of a health information manager in creating data fit for purpose,” *Health Information Management Journal*, vol. 39, no. 3, pp. 58–59, 2010.

[312] J. M. Juran and A. B. Godfrey, *The quality control process*. McGraw-Hill, 1999.

[313] N. M. Gatto, U. B. Campbell, E. Rubinstein, *et al.*, “The structured process to identify fit-for-purpose data: A data feasibility assessment framework,” *Clinical Pharmacology & Therapeutics*, vol. 111, no. 1, pp. 122–134, 2022.

[314] A. Kenyon, L. Deka, and D. Elizondo, “Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets,” *Computers & Security*, vol. 99, p. 102 022, 2020.

[315] M. Riley, K. Robinson, M. F. Kilkenny, and S. G. Leggat, “The suitability of government health information assets for secondary use in research: A fit-for-purpose analysis,” *Health Information Management Journal*, vol. 52, no. 3, pp. 157–166, 2023.

[316] S. Bosch, P. Provoost, and W. Appeltans, “Fit for purpose? identifying and resolving quality issues with marine biodiversity datasets in r,” *PeerJ Preprints*, 2018.

[317] X. Wang, Q. Duan, and M. Liang, “Understanding the process of data reuse: An extensive review,” *Journal of the Association for Information Science and Technology*, vol. 72, no. 9, pp. 1161–1182, 2021.

[318] J.-J. Boté-Vericad and M. Térmens i Graells, “Reusing data: Technical and ethical challenges,” *DESIDOC Journal of Library & Information Technology, 2019, vol. 39, num. 6, p. 329-337*, 2019.

[319] L. Bishop, “Ethical sharing and reuse of qualitative data,” *Australian Journal of Social Issues*, vol. 44, no. 3, pp. 255–272, 2009.

[320] C. S. Duke and J. H. Porter, “The ethics of data sharing and reuse in biology,” *BioScience*, vol. 63, no. 6, pp. 483–489, 2013.

[321] C. C. Kuhlthau, “Inside the search process: Information seeking from the user’s perspective,” *Journal of the American society for information science*, vol. 42, no. 5, pp. 361–371, 1991.

[322] B. Rolland and C. P. Lee, “Beyond trust and reliability: Reusing data in collaborative cancer epidemiology research,” in *Proceedings of the 2013 conference on Computer supported cooperative work*, 2013, pp. 435–444.

[323] H. Guan, S. Gautier, D. Gupta, *et al.*, “A learning-based declarative privacy-preserving framework for federated data management,” *arXiv preprint arXiv:2401.12393*, 2024.

[324] S. Wang, S. Zhuang, B. Koopman, and G. Zuccon, “Resllm: Large language models are strong resource selectors for federated search,” *arXiv preprint arXiv:2401.17645*, 2024.

[325] J. Quan, *Archival data repositories*, Accessed: 2024-06-18, 2018. [Online]. Available: `https://bids.github.io/dats/posts/2018-11-19-data-repos.html`.

[326] M. Abdin, S. A. Jacobs, A. A. Awan, *et al.*, “Phi-3 technical report: A highly capable language model locally on your phone,” *arXiv preprint arXiv:2404.14219*, 2024.

[327] M. Wu and S. Cao, “Llm-augmented retrieval: Enhancing retrieval models through language models and doc-level embedding,” *arXiv preprint arXiv:2404.05825*, 2024.

[328] S. Bufi, A. C. M. Mancino, A. Ferrara, D. Malitesta, T. Di Noia, and E. Di Sciascio, “Kguf: Simple knowledge-aware graph-based recommender with user-based semantic features filtering,” *arXiv preprint arXiv:2403.20095*, 2024.

[329] A. Kipf, R. Marcus, A. van Renen, *et al.*, *Sosd: A benchmark for learned indexes*, 2019. arXiv: `1911.13014 [cs.DB]`.

[330] R. Roberti and M. Ruthmair, “Exact methods for the traveling salesman problem with drone,” *Transportation Science*, vol. 55, no. 2, pp. 315–335, 2021.

[331] S. Ahmed, “Applications of graph coloring in modern computer science,” *International Journal of Computer and Information Technology*, vol. 3, no. 2, pp. 1–7, 2012.

[332] B. Sun, A. Zeynali, T. Li, M. Hajiesmaili, A. Wierman, and D. H. Tsang, “Competitive algorithms for the online multiple knapsack problem with application to electric vehicle charging,” *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 4, no. 3, pp. 1–32, 2020.

[333] M. Cho, “The knapsack problem and its applications to the cargo loading problem,” *Anal. Appl. Math*, vol. 13, pp. 48–63, 2019.

[334] C. N. Lintzmayer, M. H. Mulati, and A. F. da Silva, “Register allocation with graph coloring by ant colony optimization,” in *2011 30th International Conference of the Chilean Computer Science Society*, IEEE, 2011, pp. 247–255.

[335] M. Bofill, J. Coll, M. Garcia, *et al.*, “Constraint solving approaches to the business-to-business meeting scheduling problem,” *Journal of Artificial Intelligence Research*, vol. 74, pp. 263–301, 2022.