

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences

School of Electronics and Computer Science

Improvement of Biomedical Dataset Search Through the Integration of Provenance

by

Abdullah Hamed Almuntashiri

ORCID: 0000-0002-7343-6468

*A thesis for the degree of
Doctor of Philosophy*

September 2025

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

Doctor of Philosophy

Improvement of Biomedical Dataset Search Through the Integration of Provenance

by Abdullah Hamed Almontashiri

Efforts to support the application of Findable, Accessible, Interoperable, and Reusable (FAIR) principles in the biomedical research domain have led to an increase in the availability of datasets online, facilitating data exchange and reuse. This application significantly enhances research reproducibility and reduces the resources required to conduct research from scratch. As public biomedical repositories proliferate, an enormous number of datasets, encompassing various types of data, have become available to biomedical researchers. However, researchers require methods and tools that assist them in searching for and discovering relevant datasets. They still face challenges when using existing search engines, which may not be well-suited to biomedical research domains. These challenges include a lack of dataset metadata, which affects their ability to select relevant datasets.

In this research, we first sought to deepen our understanding of how biomedical researchers search for datasets and the challenges they encounter through semi-structured interviews. Based on our first study's findings, we focused on a specific challenge — the lack of provenance metadata — and its impact on the decision-making process. We then evaluated how provenance information enhances dataset search through a user study. Following this, we developed a provenance extraction tool to automatically extract provenance information from biomedical publications based on datasets and to estimate its scalability across all articles on exome sequencing experiments in PubMed. We conclude our research by evaluating the usefulness of the provenance extraction tool for dataset search through a user experience study.

The findings of this research provide a positive perspective on integrating provenance into biomedical dataset search. The results confirm the usefulness of provenance information in improving dataset search within the biomedical research domain, where the extracted information assists in enhancing decision-making and facilitates the selection of appropriate datasets.

Contents

List of Figures	ix
List of Tables	xi
Declaration of Authorship	xiii
Acknowledgements	xv
Definitions and Abbreviations	xvii
1 Introduction	1
1.1 Background	1
1.2 Motivation	3
1.3 Research Aims and Objectives	4
1.4 Research Questions	5
1.5 Thesis Structure	6
1.6 Publications	7
2 Literature Review	9
2.1 Biomedical Repositories	11
2.2 Dataset Search	12
2.2.1 Dataset Search Process	13
2.2.2 Dataset Search Taxonomies	15
2.2.3 Dataset Search in Biomedicine	19
2.2.4 Metadata	20
2.2.5 Metadata for Biomedicine	21
2.3 Provenance	23
2.3.1 Provenance Models	25
2.3.2 Provenance in Biomedicine	27
2.3.3 Workflow Visualisation	30
2.3.4 Provenance for Biomedical Dataset Search	31
2.4 Improving Provenance Annotation	31
2.4.1 Improving Other Metadata Annotations	34
2.5 Large Language Models (LLMs)	35
2.5.1 LLMs in Biomedicine	37
2.5.2 Summary	38
2.6 Prompt engineering	38
2.6.1 Zero-Shot Prompt	39

2.6.2	Few-Shot prompt	39
2.6.3	Chain-of-thought (CoT) prompt	40
2.6.4	Automatic Prompt	40
2.6.5	Prompt Engineering for Extraction	40
2.7	Summary	41
3	Methodology	43
3.1	Participant Recruitment Processes Overview	44
3.2	Methodology for Defining Biomedical Dataset Search Requirements (RQ1) . .	44
3.2.1	Semi-structured Interview design and development	45
3.2.2	Interview procedures	45
3.2.3	Interview Analysis	47
3.3	Methodology for Assessing the Effectiveness of Biomedical Provenance Infor- mation (RQ2)	47
3.3.1	Identifying critical provenance information for biomedical researchers	48
3.3.2	Survey Design and development	48
3.3.3	Dataset search tasks	49
3.3.4	Survey Validation	54
3.3.5	Survey procedures	55
3.4	Extracting provenance automatically (RQ3)	57
3.4.1	Extraction Tasks using LLMs	57
3.4.2	Provenance Extraction Prompts for Biomedical Research	58
3.4.3	Provenance Extraction Experiment	61
3.4.3.1	Collecting a set of papers from biomedical researchers . . .	61
3.4.3.2	Systematic Approach	61
3.4.3.3	Metrics for Evaluating Prompts	62
3.4.4	Scalability Experiment	64
3.4.4.1	Data collection	64
3.4.4.2	Performance Evaluation	64
3.5	Methodology for a User Experience Evaluation of Biomedical Dataset Search Enhancement Using Provenance Information (RQ4)	65
3.5.1	Interview design and development	66
3.5.2	Interview procedures	67
3.5.3	Interview Analysis	68
3.6	Ethical Approval	70
3.7	Summary	71
4	Dataset Search for Biomedical Researchers	73
4.1	Demographics	73
4.2	Findings	74
4.2.1	Data Formats and Types	74
4.2.2	Query Styles	76
4.2.3	Strategies and Places	76
4.2.4	The Importance of Metadata	77
4.2.5	Data Quality Issues	79
4.2.6	Accessibility	81
4.2.7	Dataset Assessment	82

4.2.8	Expanding the Importance of Provenance for Biomedical Researchers	83
4.3	Discussion	86
4.3.1	Comparison to Previous Work	86
4.3.2	Dataset Search Requirements for Biomedical Researchers	87
4.4	Summary	88
5	Measuring the Effectiveness of Provenance Information for Dataset Search	89
5.1	Demographics	89
5.2	Results	91
5.2.1	Comparison of usefulness of presentation options	91
5.2.1.1	Option A : Dataset metadata	91
5.2.1.2	Option B : Dataset metadata in abstract	92
5.2.1.3	Option C : Visual abstract of provenance metadata	94
5.2.1.4	Option D : Dataset provenance metadata combined with abstract	96
5.2.2	Information gained from provenance	98
5.2.3	Statistical Significance Tests	98
5.2.3.1	The differences between showing existing metadata and showing provenance information	99
5.2.3.2	Association between roles and dataset metadata	100
5.2.3.3	Association between metadata options and responses	101
5.2.3.4	Reliability tests	102
5.3	Discussion	103
5.4	Summary	104
6	Provenance Information Extractor	107
6.1	Implementation of Extractor	107
6.1.1	Output Validation	110
6.1.2	ChatGPT API	111
6.2	Provenance Extraction Prompts Experiment	113
6.3	Scalability Experiment	114
6.4	Discussion	121
6.4.1	Prompt Patterns	121
6.4.2	Scalability	123
6.5	Summary	123
7	Evaluating the Usefulness of Provenance Extraction for Dataset Search	125
7.1	Demographics	125
7.2	Qualitative Analysis Result	126
7.3	Quantitative Analysis Result	129
7.4	Discussion	132
7.5	Summary	135
8	Discussion	137
8.1	Summary	140
9	Conclusions	141
9.1	Future Work	143

Appendix A Human Study 1	145
Appendix B Survey + Statistical Analysis	153
Appendix C Human Study 3	193
References	197

List of Figures

1.1	The phases of this research consisting of 4 studies	5
2.1	The steps of dataset search process (Chapman et al., 2020)	13
2.2	The Taxonomy of Dataset Search. It includes three levels: the blue represents search style types, the orange is the systems/approaches and the green is the benchmarks.	17
2.3	Types of Metadata: Descriptive, Structural, and Administrative (Ulrich et al., 2022; Riley, 2017)	21
2.4	Terminologies for various metadata types (Mayernik, 2021)	22
2.5	An example of DataMed metadata includes several pieces of information about the dataset, such as its name, repository, identifier, and more.	23
2.6	Hierarchy of provenance metadata types (Herschel et al., 2017)	24
2.7	The core concepts of OPM (Moreau et al., 2011)	26
2.8	The core concepts of PROV: Entity, Activity and Agent (Moreau et al., 2011) .	26
2.9	ProvCaRe S3 model (Sahoo et al., 2019)	29
2.10	Gap at the intersection of the three main domains.	31
3.1	Research Questions (RQs)	43
3.2	Participant cohorts	44
3.3	Exclusion criteria flowchart for Study 1	46
3.4	An example of metadata from Datamed.	51
3.5	Abstract on Pubmed for paper (Cardinale et al., 2023)	52
3.6	Provenance information arranged as an activity-centered design.	53
3.7	Examples of the five-point Likert scale used for the usefulness assessment. . .	54
3.8	G*Power for sample size estimation.	56
3.9	Four interaction steps followed with the LLM.	63
3.10	Example of extractor evaluation by a participant.	67
3.11	Interaction workflow between users and the extractor.	68
4.1	Order of search strategies followed by our participants	78
4.2	Example of missing metadata elements	78
4.3	Example of an incomplete dataset from NNDSS (2016), with missing information in several columns, including data on Chlamydia trachomatis infection and Coccidioidomycosis.	80
4.4	Example of access restrictions from an open data portal	81
5.1	Likert-style scale for the usefulness assessment.	102
6.1	Architecture of the extractor.	108

6.2	The extractor workflow diagram.	109
6.3	Example of a limit error message.	111
6.4	Example of RDF file validation.	112
6.5	Provenance information validation.	113
6.6	Number of iterations and number of files	116
6.7	Estimated response time using the Power-Law model.	119
6.8	Cost estimation result using power-law equation.	120
6.9	Cost estimation result using polynomial equation.	121
Appendix B.1	Option A: DataMed Metadata	181
Appendix B.2	Option B: Metadata in abstract	181
Appendix B.3	Option C: Provenance metadata	182
Appendix B.4	Option D: Provenance metadata	182
Appendix B.5	Statistical tests conducted in SPSS 1	183
Appendix B.6	Statistical tests conducted in SPSS 2	183
Appendix B.7	Statistical tests conducted in SPSS 3	184
Appendix B.8	Statistical tests conducted in SPSS 4	184
Appendix B.9	Statistical tests conducted in SPSS 5	185
Appendix B.10	Statistical tests conducted in SPSS 6	185
Appendix B.11	Statistical tests conducted in SPSS 7	186
Appendix B.12	Statistical tests conducted in SPSS 8	186
Appendix B.13	Statistical tests conducted in SPSS 9	187
Appendix B.14	Statistical tests conducted in SPSS 10	187
Appendix B.15	Statistical tests conducted in SPSS 11	188
Appendix B.16	Statistical tests conducted in SPSS 12	188
Appendix B.17	Statistical tests conducted in SPSS 13	189
Appendix B.18	Statistical tests conducted in SPSS 14	189
Appendix B.19	Statistical tests conducted in SPSS 15	190
Appendix B.20	Statistical tests conducted in SPSS 16	190
Appendix B.21	Statistical tests conducted in SPSS 17	191
Appendix B.22	Statistical tests conducted in SPSS 18	191
Appendix B.23	Statistical tests conducted in SPSS 19	192
Appendix B.24	Statistical tests conducted in SPSS 20	192

List of Tables

2.1	Existing provenance models and extensions	32
2.2	Overview of studies under each approach represented in the provenance dendrite diagram.	34
2.3	Comparison of LLMs	36
3.1	List of Biomedical Articles	62
4.1	Demographic information of participants in study 1	74
4.2	Summary of provenance requirements elicited by our study.	85
4.3	Comparison of provenance findings from this study with those reported by Johns et al. (2023)	86
4.4	Comparison of findings from this study and prior research Gregory et al. (2020)	87
5.1	Demographic information of participants in Study 2	90
5.2	Helpfulness scores per presentation item for Option A, metadata only	93
5.3	Quotes of participants that chose not to download the dataset under Option A, metadata only.	94
5.4	Helpfulness scores for Option B, abstract only	94
5.5	Quotes of participants that chose not to download the dataset under Option B, abstract only.	95
5.6	Helpfulness scores for Option C, provenance information.	96
5.7	Quotes of participants that chose not to download the dataset under Option C, provenance information.	96
5.8	Helpfulness scores for Option D, provenance and abstract.	97
5.9	Quotes of participants that chose not to download the dataset under Option D, provenance and abstract.	97
5.10	Summary of provenance information gained	99
5.11	Pearson Chi-square test results for the association between roles and different dataset description approaches across tasks.	101
5.12	Fisher's Exact test results for the association between roles and different dataset description approaches across tasks	101
5.13	Statistical test results for the approaches across two tasks	102
5.14	Cronbach's alpha analysis	103
6.1	Average Percentage of Accuracy in Responses.	114
6.2	Average Precision For All Prompt Patterns.	115
6.3	Average Recall For All Prompt Patterns.	115
6.4	Iteration results	116
6.5	Equations and MAE for time estimation.	118

6.6	The observed results compared to the predicted results	119
6.7	Equations and MAE for cost estimation.	120
6.8	Comparison of observed results, power law predictions, and polynomial predictions.	121
7.1	Participants' research domains and roles	126
7.2	Participant quotations related to the themes of performance and decision-making.	127
7.3	Participant quotations related to the themes of provenance completeness and missing information.	128
7.4	Quotes of participants related to provenance accuracy.	129
7.5	Quotes of participants related to component integration.	129
7.6	Precision for provenance information components across biomedical papers	130
7.7	Recall for provenance information components across biomedical papers	131
7.8	Precision for description of provenance information components across biomedical papers.	133
7.9	Recall for description of provenance information components across biomedical papers.	134
Appendix A.1	Summary of Interview Questions and Justifications	151

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

Abdullah Hamed Almntashiri, Luis-Daniel Ibáñez, and Adriane Chapman. A taxonomy of dataset search. In *The International Conference of Advanced Computing and Informatics*, pages 562–573. Springer, 2022

Abdullah Hamed Almntashiri, Luis-Daniel Ibáñez, and Adriane Chapman. Llms for the post-hoc creation of provenance. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 562–566. IEEE, 2024

Abdullah Hamed Almntashiri, Luis-Daniel Ibáñez, and Adriane Chapman. Using llms to infer provenance information. In *Proceedings of ProvenanceWeek (PW'25)*, Berlin, Germany, June 22–27 2025. ACM. . URL <https://doi.org/10.1145/3736229.3736261>

Signed:.....

Date:.....

Acknowledgements

First and foremost, I would like to thank Allah Almighty for granting me the strength, guidance and patience to complete this journey.

I would also like to express my sincere thanks and gratitude to the many amazing people I have met and who have supported me throughout my studies.

I would like to express my deepest gratitude and appreciation to my wonderful supervisors, Prof. Adriane Chapman and Dr. Luis-Daniel Ibáñez, for their invaluable motivation and support. This journey would not have been possible without your expertise, guidance and encouragement throughout our supervision meetings. I extend my gratitude to Prof. Sarah Ennis, Dr. Tom Blount, Dr. Jacqui Ayling, and others whose contributions and consultations were instrumental to my study. My sincere gratitude is also due to Dr. Nicholas Gibbins, Dr. George Konstantinidis and Dr. Stian Soiland-Reyes for their constructive feedback and thoughtful recommendations on this PhD. I would like to thank all the participants who took part in the studies.

Additionally, my deepest gratitude goes to my family, whose unwavering support, belief in me, and constant prayers have been the foundation of this achievement and a source of strength and motivation throughout this journey. I could not have completed this without my mother's endless prayers and my father's encouragement. My dear wife, your constant patience, sacrifices, understanding, and emotional support have been my strength in achieving my dreams. I deeply appreciate all that you have done and continue to do for our family. I could not have completed this journey without your unwavering care and support. I am also deeply grateful to my son, who is the heart of my life. I hope you will forgive me for the time I have missed and for any shortcomings in fulfilling my duties towards you. Additionally, I would like to thank my siblings for their continuous support and encouragement. This acknowledgment would not be complete without thanking all my family.

To my friends who have become my family in the UK, I would like to express my heartfelt gratitude for your support during some of the most challenging times. Special thanks are due to Abdulaziz Babulghum, Murad, Baadhim, Ateeq, Omar Alharbi, Aldhahri, Alquliti, BinAfif, Sharaf and Alhashimi. Your kindness, encouragement, and presence have been truly invaluable to me.

Lastly, my heartfelt thanks extend to the Saudi Arabian government and Najran University for funding my studies. I am also deeply grateful to Dr. Belqasem Aljafari, Dr. Mohammed Aljabbar, Dr. Mohammed Alshahrani, and Mr. Abdullah Al Khalis, members of Najran University, for their support and cooperation throughout this journey. Special thanks are also due to Dr. Yasseen Almakady and Prof. Ahmed Omar Alsayed for their invaluable support.

Definitions and Abbreviations

GEO	Gene Expression Omnibus
FAIR	Findable, Accessible, Interoperable, and Reusable
IR	Information Retrieval
NLP	Natural Language Processing
DCAT	Data Catalog Vocabulary
JSON	JavaScript Object Notation
RDF	Resource Description Framework
SQL	Structured Query Language
CQL	Contextual Query Language
OPM	Open Provenance Model
PML	Proof Markup Language
NER	Named Entity Recognition
LLMs	Large Language Models
ML	Machine Learning
IE	Information Extraction
FoM	Faculty of Medicine

Chapter 1

Introduction

1.1 Background

Dataset reusability has become increasingly common among researchers in several research domains. Several motivations for reusing existing datasets have been identified in the literature. One key objective is the integration of heterogeneous datasets to enable further analysis and promote scientific advancement across a range of domains (Löffler et al., 2021). For example, Rajkomar et al. (2018) reported that the reuse of existing health datasets assisted them in developing machine learning algorithms for predictive purposes. Additionally, several government agencies and research funders promote dataset sharing and reuse, such as the U.S. National Institutes of Health (NIH) (Waithira et al., 2024; Sielemann et al., 2020). Another objective in supporting dataset reusability is to reduce the costs associated with research implementation (Weiskopf and Weng, 2013).

In the biomedical research domain, Roberts et al. (2017) confirms that researchers seek to collect, reuse and share datasets to conduct scientific experiments. This trend has contributed to the proliferation of publicly available biomedical datasets, offering substantial benefits across a range of research tasks (Cohen et al., 2017). Firstly, it facilitates the reusability of datasets generated from scientific experiments (Sarkans et al., 2018). Secondly, dataset availability offers significant potential for improving clinical research (Gierend et al., 2024). Thirdly, it can save researchers considerable time by eliminating the need to conduct studies from scratch (Zhang and Ashraf, 2023). Finally, access to datasets can play a critical role in enhancing the reproducibility of clinical trials and reducing associated costs (Valdez et al., 2017).

The reuse of existing datasets may assist in knowledge discovery and enhance the productivity and reproducibility of scientific research. The implementation of the Findable, Accessible, Interoperable, and Reusable (FAIR) principles (Jacobsen et al., 2020), along with data sharing, is essential in the biomedical research domain for several reasons: fostering an open research community; facilitating information dissemination; and improving the quality of research data (Parra-Calderón et al., 2020). Several institutes in biomedicine, including the National Sleep Research Resource (NSRR) and The Cancer Genome Atlas (TCGA), support data sharing initiatives to enhance the availability and reusability of research data (Sahoo et al., 2019). As a result, the availability of datasets in biomedical domains has expanded considerably over the past decades through biomedical repositories, enabling researchers to share, reuse, and search for datasets (Sinha et al., 2009; Wei et al., 2018; Zhang and Ashraf, 2023). Gene Expression Omnibus (GEO), DataMed, ArrayExpress, the UK Biobank, and other portals provide millions of publicly available biomedical datasets for reuse (Zhang and Ashraf, 2023). For instance, GEO¹ contains more than 7.4 million samples across over 200,000 datasets.

To facilitate the process of finding and accessing datasets, the dataset search domain has emerged. The importance of data search and discovery has been recognised across complementary disciplines (Koesten, 2019a). This recognition may strengthen interdisciplinary connections between biomedical research and the dataset search domain to better address users' needs. In particular, dataset search has the potential to promote data reuse within the biomedical domain (Chen et al., 2018).

Dataset search is the process of exploration and discovery, ultimately providing users with relevant datasets (Chapman et al., 2020). According to Wei et al. (2018); Patra et al. (2020), the large number of biomedical repositories poses a challenge for researchers in maintaining a comprehensive inventory. Additionally, the process of finding relevant datasets can be complicated by the volume and complexity of biomedical data (Chen et al., 2018). Therefore, the development of specialised search engines in the biomedical research domain has become increasingly recognised, as they have the potential to improve dataset searchability and reusability (Roberts et al., 2017). For instance, DataMed² is a widely used open-source data discovery system dedicated to biomedical datasets. It provides access to 49 repositories, 20 data types, and over one million datasets.

While the availability of biomedical repositories and search engines have facilitated data sharing and reusing, it has also introduced challenges in data search and retrieval, requiring more advanced search methodologies. According to Paton et al. (2023), issues related to the effective

¹<https://www.ncbi.nlm.nih.gov/geo/>

²<https://datamed.org/> (accessed on 30 June 2025)

retrieval of available datasets across the web have emerged. Although there have been advancements in Information Retrieval (IR) and Natural Language Processing (NLP) techniques, dataset search remains less developed than other search verticals (Kacprzak et al., 2018), which focuses on particular subjects and aggregates data from different sources (Lewandowski, 2023). The field of dataset search is still novel, and to date, limited research has been conducted on this domain, whether these datasets exist on the web or within specialised portals (Koesten, 2019a). This limitation is not restricted to a specific domain but extends across multiple fields, including data science (Koutras et al., 2021), software engineering (Wang et al., 2019b), geographical information science (Yang et al., 2024), and biomedicine (Waldrop et al., 2021).

Wei et al. (2018) indicate new obstacles and challenges in the domain of biomedical information retrieval. Similarly, Wang et al. (2019c) confirm that searching, curating, and annotating publicly available gene expression datasets online for reuse in further research presents significant challenges.

1.2 Motivation

With the continuous advancement in dataset generation and production in biomedical research, the discovery and search for available datasets has become increasingly essential, facilitating the sharing and reuse of these datasets. This progress supports the implementation of FAIR principles, which prompts several functions such as research reproducibility and data reuse. However, Hughes et al. (2023) confirm that biomedical researchers face challenges in achieving the FAIR principles in biomedical dataset search, most notably a lack of metadata collection and accessibility. Similarly, Waldrop et al. (2022) highlight the difficulties of searching for and identifying appropriate biomedical datasets among those available. Moreover, there is a clear need to develop new techniques and methods to assist researchers in searching for relevant biomedical datasets within data repositories (Löffler et al., 2021). For instance, publicly available gene expression datasets face challenges related to searching, curation, and annotation, which impact the reuse in further research (Wang et al., 2019c).

These challenges in biomedical dataset search are connected to several key aspects:

Firstly, there is a lack of understanding of how researchers approach their current data-searching tasks and the role that dedicated data search portals play in their information-seeking processes (Krämer et al., 2021). Understanding the behaviours of biomedical researchers during dataset search remains an ongoing challenge (Wei et al., 2018). Addressing this aspect may support

technical specialists in developing or improving biomedical dataset search techniques and systems.

Secondly, the diversity of biomedical data types, along with their formats and presentation, poses significant challenges in dataset search (Wei et al., 2018; Bouadjene and Verspoor, 2017; Wang et al., 2019c). Roberts et al. (2017) mentioned that finding relevant data is highly challenging because of the diversity of data types linked to a dataset, as well as the growing volume of sources.

Thirdly, the metadata provided is insufficient to cover all essential details linked to datasets (Dixit et al., 2018). Löffler et al. (2021) confirm that existing metadata poorly reflect user requirements, leading to difficulties in retrieving relevant biodiversity research datasets. In addition, the lack of metadata in data lakes and repositories complicates the process of integrating separate datasets into a single searchable index (Tsung et al., 2023). This challenge also impacts the reproducibility of scientific experiments (Gierend et al., 2023). For example, the metadata within GEO does not include vocabularies or data to define key biological information, such as cell lines and cell types (Wang et al., 2019c), which are essential for experimental reproducibility. As a result, dataset searchers are often forced to rely on literature searches to investigate related publications and gather additional information (Krämer et al., 2021; Xiao et al., 2019).

At a higher level of public data repositories, some dataset search systems do not offer tailored support biomedical datasets. For instance, Google Dataset Search encounters obstacles in meeting the requirements of biomedical researchers due to the limitations of a broad schema in supporting biological data, the diversity of coding and units in experimental data and the variability in data formatting in biomedicine (Wang et al., 2019c).

1.3 Research Aims and Objectives

This research aims to expand our understanding of biomedical dataset search, explore current challenges, and propose potential improvements to support biomedical researchers in their research journeys. Figure 1.1 illustrates the phases of our research. The first study explores the current dataset search process used by biomedical researchers and identifies the difficulties they encounter. Following that, a user study was conducted to assess the effectiveness of the availability of provenance information in improving biomedical dataset search. Based on this study, we then developed a provenance extraction tool designed to extract provenance information from papers published on datasets. These papers describe biomedical experiments conducted to generate the data within the dataset. Finally, we enhanced this research by measuring the extractor

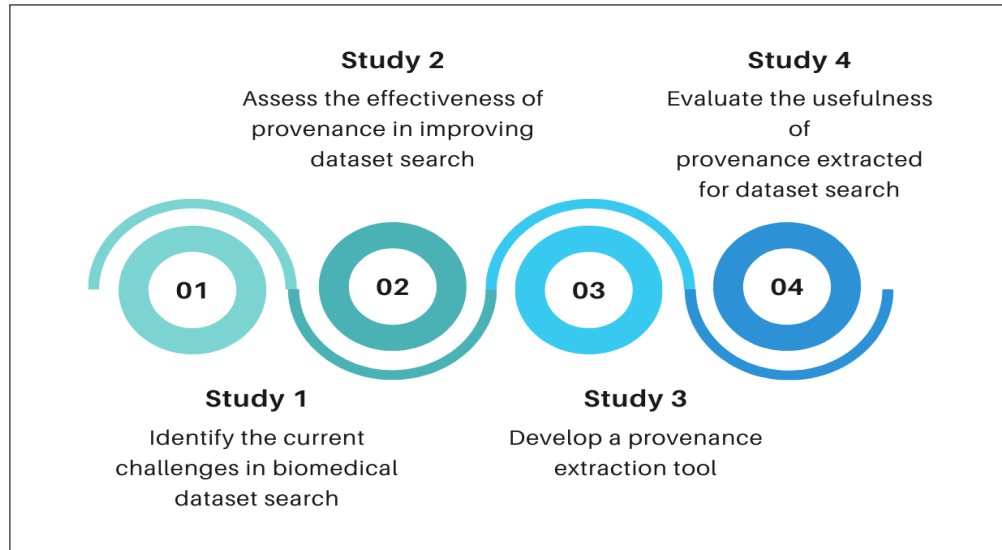


FIGURE 1.1: The phases of this research consisting of 4 studies

across a wide range of biomedical research publications to estimate its scalability and conducted a user experience evaluation study.

The research objectives of this work are:

1. Understand the current dataset search practices in the biomedical domain.
2. Identify the challenges in how dataset search is conducted.
3. Assess the effectiveness of including provenance information in biomedical dataset search.
4. Develop a provenance extraction tool to enhance and facilitate the biomedical dataset search process.
5. Evaluate the usefulness of the provenance extraction tool.

1.4 Research Questions

To achieve our objectives, we pose the following research questions:

- RQ1: What are biomedical researchers' dataset search requirements?
- RQ2: What provenance is needed by biomedical researchers for dataset search?
- RQ3: How and with what accuracy can we infer provenance information from textual descriptions in biomedical research papers?

- RQ4: To what extent does the extracted data provenance from the biomedical publications help biomedical researchers do dataset search?

1.5 Thesis Structure

The following chapters of this thesis is structured as follows:

Chapter 2 provides background on dataset search, its process, and related work, including existing taxonomies and metadata. We also present a brief overview of provenance, its models, and its applications in biomedicine. Additionally, we introduce LLMs, their use in biomedicine and the concept of prompt engineering in LLMs and highlight its importance for generation of text.

Chapter 3 presents the research methodologies used to address the research questions. It includes the methods, tools, and techniques employed in all studies.

Chapter 4 presents the findings of our first study, which aims to identify the current dataset search techniques used by biomedical researchers and the challenges they encounter. Additionally, we outline and discuss the essential requirements for improving dataset search in biomedical research.

Chapter 5 displays the findings of our second study, which evaluates the effectiveness of provenance information in enhancing biomedical dataset search. Moreover, it identifies the key provenance elements required when searching for biomedical datasets.

Chapter 6 focuses on the implementation of an LLM based provenance extractor for biomedical articles. It presents the architecture, components, and prompts used in the development of this tool. Additionally, we explain and discuss a scalability experiment conducted to estimate the expansion of the extractor's coverage across a wide range of biomedical publications.

Chapter 7 presents the final study, which aims to evaluate the extractor's performance through a user experience assessment. It includes the results of both qualitative and quantitative analyses.

Chapter 8 provides an overall discussion of all the studies in this thesis and summarises how the results connect to the wider literature.

Chapter 9 concludes and summarises our work. Additionally, it provides suggestions for potential future research.

1.6 Publications

- Almntashiri, A.H., Ibáñez, L.D., & Chapman, A. 2023. A Taxonomy of Dataset Search. In: Saeed, F., Mohammed, F., Mohammed, E., Al-Hadhrami, T., Al-Sarem, M. (eds) *Advances on Intelligent Computing and Data Science. ICACIn 2022. Lecture Notes on Data Engineering and Communications Technologies*, vol 179. Springer, Cham. https://doi.org/10.1007/978-3-031-36258-3_50
- Almntashiri, A.H., Ibáñez, L.D. and Chapman, A., 2024, July. LLMs for the post-hoc creation of provenance. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (pp. 562-566). IEEE.
- Almntashiri, A.H., Ibáñez, L.D., & Chapman, A. 2025. Using LLMs to Infer Provenance Information. In *ProvenanceWeek (PW'25)*, June 22–27, 2025, Berlin, Germany. ACM. <https://doi.org/10.1145/3736229.3736261>

Chapter 2

Literature Review

This chapter provides a comprehensive background and foundational knowledge on all subjects relevant to this research. We present a brief description of the concept of a “dataset”, including its definition, significance, and common formats. Section 2.1 offers an overview of dataset repositories and portals and presents several examples of biomedical dataset repositories. In Section 2.2, we provide a foundational overview of dataset search, including its definitions, processes, procedures, and existing dataset search taxonomies. Additionally, we review the use of dataset search in biomedicine and the challenges faced by biomedical researchers. Furthermore, we explore the use of metadata in dataset search as a method for discovering retrieved datasets and selecting those that best meet users’ needs, as well as its role in biomedical research domain. In Section 2.3, we highlight a type of metadata (provenance), its usage in various domains and tasks. Additionally, we explore popular provenance models, including those used in biomedicine. Section 2.4 presents topics related to improving provenance annotation, including methods, classification approaches and other forms of metadata annotation. In Section 2.5, we provide a brief explanation of LLMs and their applications in biomedical domains. Finally, in Section 2.6, we discuss prompt engineering as a technique to enhance the output of LLMs and its use in extraction tasks.

Before describing the dataset, we need to understand the concept of data. Data are “the raw observations about the world collected by scientists and others, with a minimum of contextual interpretation” (Zins, 2007, p.484). In addition, data are “sets of characters, symbols, numbers, and audio/visual bits that are represented and/or encountered in rawforms” (Zins, 2007, p.485). Data can be presented in various forms, such as images, graphs, or alphanumeric characters like letters and numbers. To be included in a dataset, data should be organised, structured, and properly formatted.

A dataset has been defined by various communities and individuals, with definitions varying depending on the specific interests and focus of each community. The term “dataset” may refer to analogue or physical data collections (such as handwritten notes or biological samples) (Borgman, 2015). Additionally, the same term can also refer to a dataset as a collection of documents encompassing various data types, such as images, videos, tabular data, and structured data — typically in digital form (Nguyen et al., 2020). Chapman et al. (2020) describe a dataset as an organised set of observations that can be utilised for targeted purposes. The Statistical Data and Metadata Exchange (SDMX) initiative offers a similar definition, describing a dataset as “a collection of related observations, organized according to a predefined structure” (Chapman et al., 2020, p.252). Similarly, Maali et al. (2014) highlight the Data Catalog Vocabulary (DCAT), which defines a dataset as a set of data curated or issued by an individual agent, designed to be accessible and downloadable in various formats. Koesten (2019a, p.11) define a dataset as “structured information collected by an individual or organisation, distributed in a standard format”.

In this thesis, we adopt the common understanding of a dataset as a digital representation encoded in structured or semi-structured formats. Various standard formats are widely used for storing and sharing such data. For instance, commonly used formats include Comma-Separated Values (CSV), Resource Description Framework (RDF), and JavaScript Object Notation (JSON). According to Koesten (2019a), tools for discovering data often support multiple formats, such as CSV, XML, and HTML, for downloading data files.

With the growing volume of datasets, storing large amounts of data locally has become increasingly challenging. The expansion of data availability has led to the development of open data platforms, repositories, portals and marketplaces. Data platforms provide the infrastructure for data ecosystems, supporting users in interacting with open data contained within datasets. Such interactions include searching, sharing, processing, and analysing related datasets (Sennaik et al., 2017). An example of such a platform is Google BigQuery¹. Another example is OpenDataSoft², which hosts over 34,000 datasets and has recorded approximately 855 million downloads.

Open data portals are web-based interfaces that provide access to datasets offered by governments, institutions, or organisations (Thorsby et al., 2017). Numerous open data portals, such as the UCI Repository³ and the UK’s Data Service⁴, have been established to make datasets publicly accessible (Kassen, 2013).

¹<https://cloud.google.com/bigquery>

²<https://data.opendatasoft.com/pages/home/>

³<https://archive.ics.uci.edu/>

⁴<https://ukdataservice.ac.uk/>

Several data repositories have emerged — storage systems or archives in which datasets are deposited, curated, and preserved for long-term access and reuse (Duerr, 2014). Examples of generic repository software include Dataverse⁵, Zenodo⁶, Figshare⁷, and CKAN⁸. DataPlanet⁹ is a repository that offers around 13.5 billion datasets provided by more than 90 data providers across 16 major domains. Thøgersen and Borlund (2022); Marcial and Hemminger (2010) highlight the growing availability of datasets in scientific repositories, such as Elsevie¹⁰, which play a critical role in supporting various research fields (Altman et al., 2015).

In addition, several discipline-oriented data repositories, portals, and marketplaces provide access to specialised datasets. Grossman et al. (2006) note that many health institutions aim to share datasets to support treatment development, including medical images, clinical data and omics data. Beyond public portals and repositories, the emergence of data markets has further elevated datasets as a significant source of trade (Grubenmann et al., 2018). Furthermore, the European Data Portal¹¹ publishes approximately 1,781,000 datasets from 35 EU countries, spanning over 191 catalogues. Additionally, FAIRsharing¹² includes more than 2,000 registered databases, many of which are domain-specific and focused on the life sciences.

2.1 Biomedical Repositories

A large number of scientific data platforms are at the core of this research and have been developed for biomedical data, including GEO, DataMed, ArrayExpress¹³, and the UK Biobank¹⁴. A common gene expression data repository provides genome methylation, chromatin structure and functional genomics data to the public (Patra et al., 2020). GEO was established in 2000 with the support of the US National Center for Biotechnology Information (NCBI) to aid gene expression studies (Clough and Barrett, 2016). This repository contains a vast amount of data, with over 200,000 dataset series available, including more than 7.4 million samples. This repository provides several services for biomedical researchers, including querying using keywords, visualising, and analysing data.

⁵<https://dataverse.org/>

⁶<https://zenodo.org/>

⁷<https://figshare.com/>

⁸<https://ckan.org/>

⁹<https://dataplanet.sagepub.com>

¹⁰<https://elsevier.digitalcommonsdata.com/research-data/>

¹¹<https://www.europeandataportal.eu>

¹²<https://fairsharing.org/search?fairsharingRegistry=Database>

¹³<https://www.ebi.ac.uk/biostudies/arrayexpress>

¹⁴<https://www.ukbiobank.ac.uk/>

Another common open-source data discovery system for biomedical datasets is DataMed. It provides access to 49 repositories and 20 data types, including clinical trial data, and includes over 1.2 million datasets. The primary search technique is keyword-based, using a Google-like search box. Additionally, it offers advanced search tools, including Boolean operators (Chen et al., 2018).

The Medical Research Council and Wellcome Trust developed UK Biobank in 2012 as a large-scale, open-access database and research resource for discovering medical datasets (Collins, 2012). The repository consists of various data types, including genetic data and medical images (Biobank, 2014). This data was collected from 500,000 UK Biobank participants.

BioStudies¹⁵ is a public repository that organises data from biological studies. It provides genomics datasets generated from genomics experiments for reuse by the research community (Sarkans et al., 2018). A keyword-based search style is used to provide metadata descriptions of the datasets. This repository contains more than 2.4 million biological studies, including datasets.

Technological developments over the past decades have enabled many different methods for searching for datasets. However, issues related to effectively retrieving available datasets across the web have emerged (Paton et al., 2023).

2.2 Dataset Search

The domain of dataset search and discovery has emerged as an integral area within several complementary disciplines (Koesten, 2019b). The importance of this domain has increased among researchers, as they require access to datasets (Akujuobi and Zhang, 2017).

However, as this research direction is still in its infancy, the development of dataset search techniques remains limited. There is also a lack of supporting tools to help users find datasets that meet specific requirements (Wang et al., 2019b). The difficulty of effectively locating datasets across the web has become a major issue, leading dataset search to be a prominent topic in the information retrieval domain (Paton et al., 2023). Therefore, various efforts by researchers and scientists have been made to define the fundamental pillars of this domain. Several definitions outlining the process of dataset search have been proposed as its foundational basis.

- Chapman et al. (2020) define the domain of dataset search as the process of exploring and discovering datasets in order to provide them to end users seeking such datasets.

¹⁵<https://www.ebi.ac.uk/biostudies/>

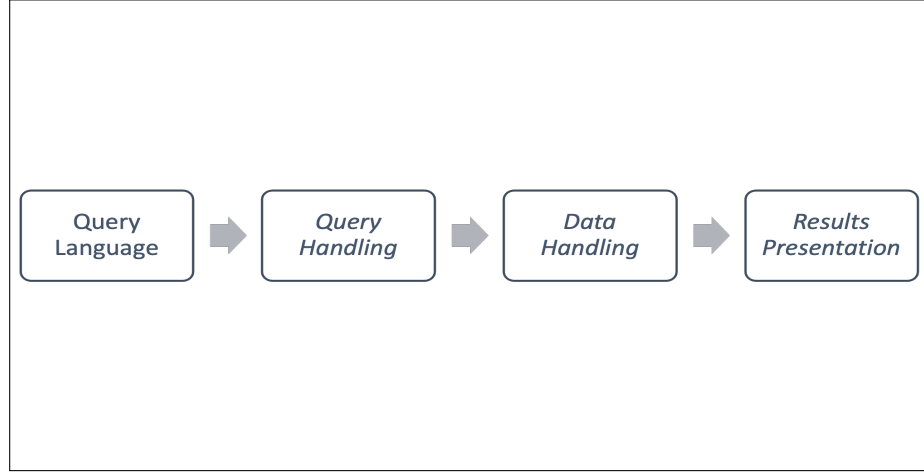


FIGURE 2.1: The steps of dataset search process (Chapman et al., 2020)

- Another definition is “the process of returning relevant RDF datasets” (Kunze and Auer, 2013, p.1).
 - Koesten (2019a) describes this domain from the perspective of data users as the process of searching for particular data within datasets.
 - A definition given by Kato et al. (2021) refers to dataset search as the function of receiving a user prompt or query, which is processed by a system to produce a ranked collection of datasets.
 - Chen (2022, p.25) defined “dataset search” as the process “to rank a set of datasets $D = \{D_1, D_2, \dots, D_n\}$ in descending order of their relevance scores with respect to q ”.
- The above definitions are explicit definitions of the term dataset search. Many other definitions are available from various communities, such as the Information Retrieval Community or Database Community, which could be used to define dataset search due to the convergence of these communities and the use of similar terms.
- According to Koesten (2019a), the term of “dataset search” and the term “data discovery” are predominantly utilised interchangeably.

2.2.1 Dataset Search Process

To better understand the dataset search process, Chapman et al. (2020) provides a high-level overview of the process and its fundamental steps in their survey. This process begins with submitting a query and ends with returning a ranked list of datasets (Chapman et al., 2020; Kato et al., 2021). Figure 2.1 illustrates these steps of the search process.

The first step of this process is submitting a query, which involves receiving and processing a user query. It is important to understand how data are converted into input queries (Chen et al., 2019). Various types of queries exist, including keyword-based queries (Ibáñez and Simperl, 2022; Hulsebos et al., 2024) and Contextual Query Language (CQL) expressions (Chapman et al., 2020). Additionally, a dataset sample can be submitted as a query to retrieve relevant datasets (Koutras et al., 2021).

The second step is query handling. This step is a critical task involving the processing of the received query to retrieve the desired results. Several existing dataset search systems adopt a metadata-based approach, relying on the similarity between the metadata and the user query (Hulsebos et al., 2024). However, low quality metadata impact the process of dataset discovery and consumption (Chapman et al., 2020; Löffler et al., 2021). Several studies argue that standard metadata does not provide sufficient information to enable users to reuse datasets (Nguyen et al., 2015; Koesten et al., 2020; Löffler et al., 2021; Sostek et al., 2024). Therefore, several studies have discussed certain surrogates, for example, textual or visual ones, that aim to assist people in determining which documents are relevant (Wiggins et al., 2018; Koesten et al., 2020). Given the importance of dataset search, several techniques have been developed to process user queries. For example, Thomas et al. (2015) utilise the type of data and column information to link query columns to underlying table columns. Another approach developed that allows keyword queries over columns (Pimplikar and Sarawagi, 2012). Additionally, Zhang and Balog (2018) proposed a method for searching based on a keyword query over a table.

The third step is data handling. This step involves preparing the results retrieved from the search process by providing a summary of each dataset, referred to as metadata, to help users make informed decisions. According to (Chapman et al., 2020; Löffler et al., 2021; Hulsebos et al., 2024), dataset publishers provide metadata to meet users' needs. This metadata is structured in a uniform format to ensure consistency in data types. To facilitate the indexing of this metadata, several vocabularies have been developed, including Schema.org¹⁶. Brickley et al. (2019) highlight that data publishers use Schema.org or W3C DCAT markup to characterise the attributes of a dataset search. Data handling can also enhance the effectiveness of a search, with some studies focusing on quality and entity resolution (Chapman et al., 2020). For example, one proposed method for addressing the issue of metadata quality involves combining feeds from identified entities, which can be used to cross-validate the metadata (Heyvaert et al., 2015). Another approach involves investigating techniques to detect coverage and bias in a dataset, which may affect algorithms using the dataset as input, a process known as schema matching (Asudeh et al., 2019). Moreover, a table similarity approach for table extension is employed in

¹⁶<https://schema.org/>

constructive dataset search, such as in WikiTables. Additional techniques, such as summarisation and annotation, are used to enhance the effectiveness of a search (Chapman et al., 2020).

The final step is the presentation of results, which aims to present the search outcomes in a structured manner. One commonly used presentation method in many data portals is the traditional ten blue links paradigm. Several Search Engine Results Pages (SERPs) used for dataset search also follow the traditional ten blue links paradigm (Chapman et al., 2020; Ibáñez and Simperl, 2022). Additionally, several frameworks are proposed for data search and exploration, including TableLens (Pirolli and Rao, 1996), DataLens (Liu and Jagadish, 2009), and the Relation Browser, which is used for sense making with statistical data (Marchionini et al., 2005). In addition, the Google Dataset Search uses a list, which is a traditional way to present results, to display the results (Chapman et al., 2020; Sostek et al., 2024).

2.2.2 Dataset Search Taxonomies

Several categorisation efforts have been undertaken to classify existing search techniques and systems. A taxonomy of web search queries was developed by (Broder, 2002), focusing on user needs. In this study, user queries are categorised into three types: First, navigational queries, which are used to find targeted web pages; Second, transactional queries, which are utilised to conducted actions on the web, such as online shopping; Third, informational queries, which aim to access specific information. A limitation of this taxonomy is that it focuses solely on web search activities and may not be directly applicable to the contexts of dataset search.

McDonnell and Shiri (2011) developed a taxonomy aimed at understanding the variations among approaches, systems, and techniques in social search. This taxonomy is divided into five categories: First, collaboration type — includes synchronous and asynchronous collaboration of user interactions during data search; Second, it distinguishes collaboration into implicit and explicit; Third, search target — focuses on targets of search, including finding people - centric systems and finding information-centric systems; Fourth, finding approaches — identified as goal-oriented search approaches and exploratory finding approaches; Finally, search result presentation — classified into sense-making results and ranking based on relevance of results. However, this taxonomy was exclusively focused on social search.

Another taxonomy was developed with a focus on data-centric tasks (Koesten, 2019b). This taxonomy centres on the activities and interactions of users with data. It is categorised into two main types: process-oriented and goal-oriented tasks. The first type focuses on tasks that are considered transformative in nature from the users' perspective. This type involves several tasks,

including working with data using various tools, integrating different datasets, utilising data in ML processes, and producing new data. The latter type includes tasks that involve using data to accomplish broader goals. Examples include answering specific questions, comparing data or datasets, and engaging in similar activities. However, this taxonomy may not be sufficient to cover the full scope of the dataset search domain, as it focuses solely on users' interactions with data.

Although the aforementioned efforts classify the searching process from different dimensions, these works may not be entirely sufficient for the dataset search process, as they do not encompass search style techniques, which are an integral part of the search process.

As part of this project, we have therefore developed a **taxonomy of dataset search** ([Almuntashiri et al., 2022](#)). During the search process to construct this taxonomy, we used four main sources: Google Scholar as a search engine; Science Direct as a scientific platform; IEEEExplore Digital Library; ACM Digital Library as digital libraries. We did an initial selection based on title, year and accessibility. In addition, we used the following keywords to find the articles: "Dataset Search", "Data Search", "Data Retrieval", "Structured Data Queries", "Dataset Search Queries", "Table Queries", "How to Search for Datasets" etc. We selected 108 papers by conducting a comprehensive scanning of the title, abstract, introduction and conclusion. Thereafter, the following exclusion criteria were applied in this study: a) Studies that do not focus on searching dataset/structured data directly; b) Studies that do not clearly explain the search style used; c) Studies that lack a full-text version at the source; d) Studies that do not include the used methodology to develop their contributions. This led to approximately 31 papers of algorithms and benchmarks that focus on the dataset search directly. We sought to discover the main differences between those papers.

By reading the selected papers to build this taxonomy, we discovered that one of the main differences between the dataset search algorithms is the search style, which is the input form used when searching for a dataset. There are many basic components of searching for datasets, such as the input query and the result presentation. Since the input query enables users to express their requirements for datasets, we decided to choose the search style to be the main pillar of our taxonomy. Figure 2.2 illustrates the design of the taxonomy of dataset search as well as the algorithms, systems and benchmarks under each search style. Four main search styles were discovered, which are keyword search, search by structured query language, search by using schema matching and search by using tables. These four styles are the ones used to search for datasets/structured data in the dataset search papers.

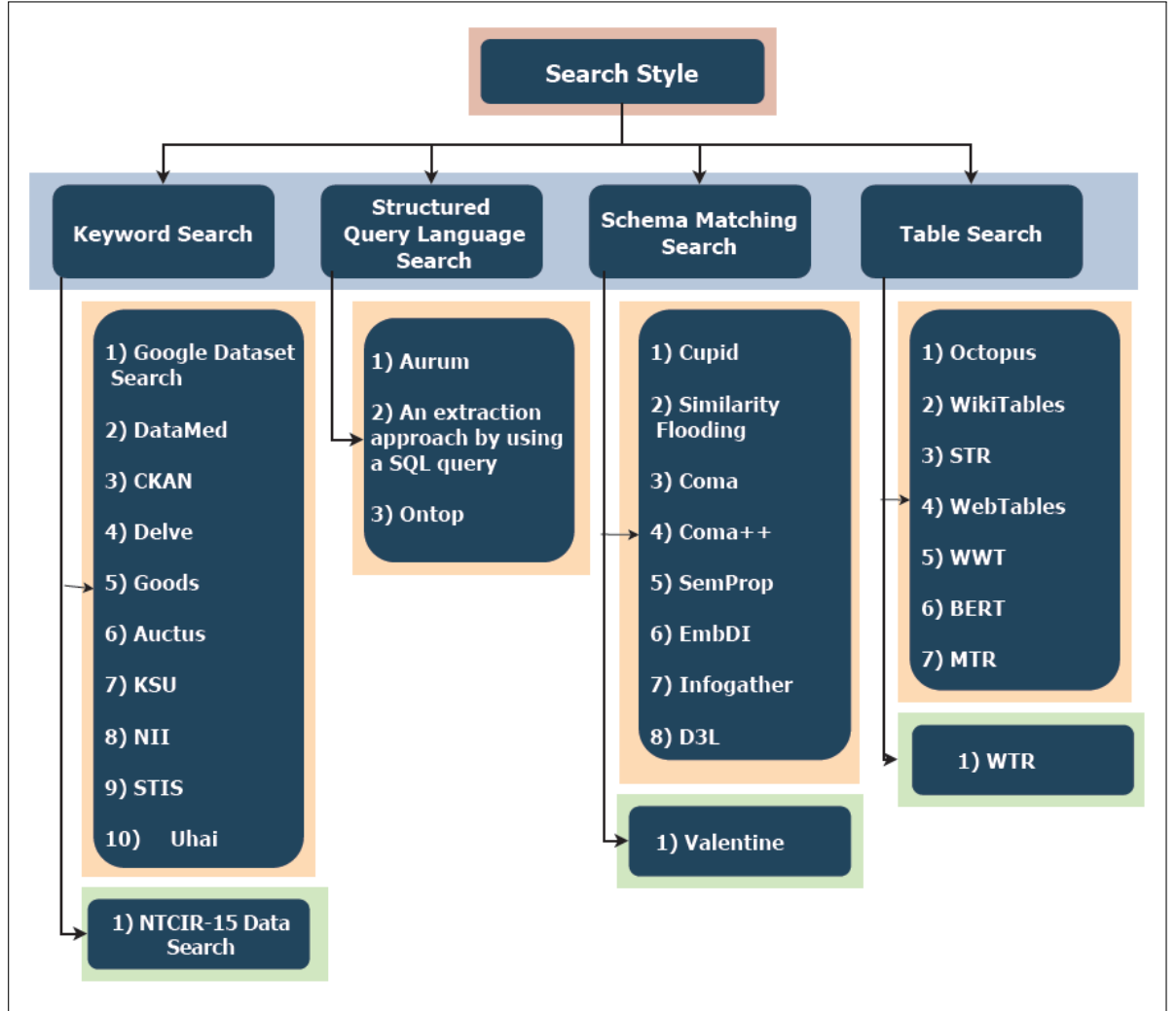


FIGURE 2.2: The Taxonomy of Dataset Search. It includes three levels: the blue represents search style types, the orange is the systems/approaches and the green is the benchmarks.

The first category is the keyword search, which includes systems and tools that utilise keywords to search for published datasets over metadata. This approach enables users to input a query related to their information needs. The second category is the structured query style, which employs structured query languages such as CQL (Contextual Query Language) and SQL (Structured Query Language). The third category is the schema matching style, which allows users to submit a dataset as input to retrieve related datasets. The final category is the table search style, which involves processing a search query and providing a list of tables, where the input query consists of either a collection of keywords or a table.

Due to the widespread distribution of existing datasets and their usage across various domains, several search styles have been developed to facilitate dataset discovery and retrieval. However, each style typically makes fundamental assumptions about the specific domains or user groups it targets. Below, we elaborate on the factors that influence the choice of search style.

Discipline. Firstly, search styles often vary across disciplines and reflect the differing extent of users' needs in dataset discovery. For example, DataMed, a biomedical data discovery system, has been designed for the biomedical research community. Consequently, it incorporates techniques that allow expert users to specify particular characteristics and supports the use of Boolean operators for specialised queries.

User Knowledge. Secondly, the employed search styles can depend heavily on the user's level of experience in dataset discovery. For instance, communities that rely on keyword-based search differ significantly from those using schema-matching approaches. While the biomedical community is data-literate, its members are generally not considered power users in database management, information retrieval, or dataset search. Thus, they predominantly utilise keyword-based searches. In contrast, the database research community — characterised by expert users — routinely employs schema matching to integrate and augment datasets. Accordingly, search styles diverge based on the intended user population for whom the tools are designed.

Data Generation, Ownership, and Access Restrictions. The origin of the dataset, the conditions under which it was created, and the associated restrictions significantly influence the appropriate search style. For example, biomedical researchers may be required to publish data to comply with grant mandates, but they are often not incentivised by personal success metrics to share data. Conversely, several European governments have adopted openness of government data as a success indicator, thereby encouraging data publication. Thus, the entity responsible for data generation — along with the level of resourcing and accessibility — affects the selection of appropriate search styles.

Task Complexity. Finally, the complexity of the task for which the dataset is intended also influences the chosen search style. This includes the required data type, such as structured tables or unstructured content. Task complexity is closely linked to user expertise and the availability of tools tailored to the needs of a given community.

From a broader perspective, the four dataset search styles identified in our survey can be positioned within the framework of classical IR techniques (Manning, 2008). Keyword search aligns closely with ad hoc information retrieval, where users submit queries and the system retrieves results ranked by estimated relevance (Kuo et al., 2024). Structured query languages, which represent our second search style, can be considered analogous to Boolean retrieval technique in IR (Salton et al., 1983), and can be enhanced through an explicit query structure. Schema matching and table search approaches share strong similarities with query-by-example (Zloof, 1977), in which the input query—such as a dataset or a table—is used to retrieve similar results. Furthermore, IR interaction techniques such as query expansion — adding related terms, synonyms, or schema

elements to improve the results — and relevance feedback — refining queries based on user judgments of result relevance — could be applicable to all four search styles, although they remain unexplored in the dataset search literature.

2.2.3 Dataset Search in Biomedicine

With technological advancements, providing data, tools, and software have become integral to contemporary biomedical domains and behavioural research (Ohno-Machado et al., 2017). The volume of datasets in biomedical domains, including molecular biology, has increased exponentially over the past decades (Sinha et al., 2009; Wei et al., 2018; Zhang and Ashraf, 2023). Mervis (2012) also emphasised that research in this domain produces a huge volume of datasets, which are stored in various formats across different sites. For example, several repositories and open data portals, such as GEO, DataMed, ArrayExpress and the UK Biobank, focus on providing researchers and users with biomedical datasets (Zhang and Ashraf, 2023). These repositories offer large biomedical datasets containing various types of data, such as omics data (Bouadjene and Verspoor, 2017). Through such platforms, biomedical data can be aggregated, searched, discovered and retrieved according to the needs of biomedical researchers.

However, it can be challenging for biomedical researchers to be aware of all public biomedical repositories that allow dataset search (Wei et al., 2018; Patra et al., 2020). Thus, the importance of search engines in the biomedical research domain has become increasingly recognised. Emerging concerns in this context include experimental reproducibility, dataset searchability and dataset reusability (Roberts et al., 2017). Wei et al. (2018) highlights several data retrieval challenges faced by the biomedical research domain, including the search behaviours of researchers when searching for datasets. Similarly, Waldrop et al. (2022) note that researchers struggle to search for and identify appropriate datasets amid the growing volume of biomedical data.

The availability and enhancement of dataset search capabilities serve several purposes: ensuring reusability and experiment replication (Patra et al., 2020; Zhang and Ashraf, 2023), addressing biomedical data scarcity and supporting research progress (Ohno-Machado et al., 2017; Waldrop et al., 2021), saving researchers' time (Zhang and Ashraf, 2023), and adhering to the “FAIR” data principles (Wilkinson et al., 2016). Additionally, access details for biomedical datasets are crucial for promoting the reproducibility of research findings (Alsheikh-Ali et al., 2011; Roberts et al., 2017). Furthermore, the availability of biomedical datasets is an integral requirement for publication conditions in several journals (Bishop et al., 2015). Dataset search is the initial step of curation, interoperability and quality of data (Hughes et al., 2023).

Although several IR and NLP techniques are employed to utilise these datasets (Benson et al., 2012; Waldrop et al., 2021), the biomedical information retrieval domain constantly faces new obstacles and challenges (Wei et al., 2018). Wang et al. (2019c) confirm the challenges associated with searching, curating, and annotating publicly available gene expression datasets online for reuse in further research. Additionally, Hughes et al. (2023) refer to biomedical researchers struggling to find, access, interoperate, and reuse (FAIR) datasets.

These challenges of biomedical datasets search are summarised in several dimensions: the presence of heterogeneous data (Bouadjenek and Verspoor, 2017; Wei et al., 2018; Wang et al., 2019c), insufficient metadata for available datasets (Bouadjenek and Verspoor, 2017; Wang et al., 2019c; Löffler et al., 2023), issues with dataset ranking (Teodoro et al., 2017) and lack of data quality (Ohno-Machado et al., 2017).

2.2.4 Metadata

Over the past few decades, metadata has been discussed across several domains to support data searchers. Several studies have mentioned that most biomedical dataset search systems, such as DataMed (Chen et al., 2018), are constructed based on the metadata of datasets (Hendler et al., 2012; Kassen, 2013; Wei et al., 2018). Before reviewing metadata in biomedicine, this section aims to provide an overview of the foundational concepts of metadata. Despite variations in how metadata is defined, there is broad agreement on the same underlying concept: “data about data” (Greenberg, 2005) or “structured data about data” (Duval et al., 2002).

A recent systematic review by Ulrich et al. (2022) discussed the concept of metadata, describing it as a detailed description of data. Several scientific communities have adopted and studied the use of metadata, including computer science, library and information science (Greenberg, 2005), and medical informatics (Ulrich et al., 2022). Additionally, metadata have been used in open data portals, public repositories and search engines (Hendler et al., 2012; Kassen, 2013) to enable users to search for datasets based on metadata. There are several examples of metadata use in dataset search tools, such as Google Dataset Search (Brickley et al., 2019) and DataMed (Chen et al., 2018).

Several types of metadata have been discussed in various studies (Ulrich et al., 2022; Riley, 2017), including descriptive metadata, structural metadata, and administrative metadata, as shown in Figure 2.3. Hartig and Zhao (2010); Muniswamy-Reddy and Seltzer (2010); Mayernik (2021); Wittner et al. (2023) confirm that provenance metadata is a common type of metadata that describes the history of data. Additionally, Mayernik (2021) states that there are 19 types of

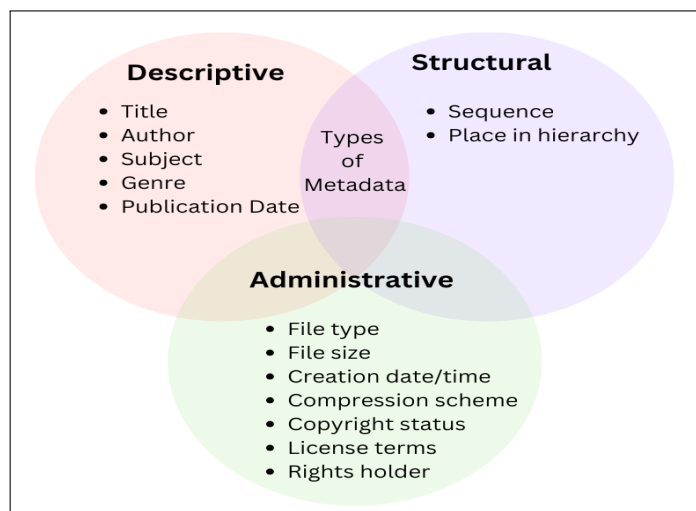


FIGURE 2.3: Types of Metadata: Descriptive, Structural, and Administrative (Ulrich et al., 2022; Riley, 2017)

metadata, including provenance metadata, each with its own concept and motivation. Figure 2.4 presents The terminologies associated with different metadata types.

The standards of metadata are categorised into three main types in the systematic review by (Ulrich et al., 2022). The first type comprises various structural standards, such as ISO 11179, OMOP, ONIX and Dublin Core. The second includes different technical standards, including XML and RDF. The third one consists of several types of semantic standards, such as UMLS and RxNorm. These types of metadata are employed for several key tasks, including dataset definition, secondary data use, information retrieval tasks and data integration tasks (Ulrich et al., 2022).

2.2.5 Metadata for Biomedicine

As stated earlier, several biomedical dataset discovery tools, such as DataMed, were built on the metadata of datasets rather than retrieving data based on their actual contents. Metadata of scientific experiments, specially in biomedicine, has been studied for a long time. Providing a high quality of metadata is integral in seaching and reusing scientific datasets located in online repositories (Gonçalves and Musen, 2019). Several contributions have been made to support researchers in using biomedical data. For example, Wang et al. (2019c) developed dedicated metadata for gene expression datasets, aiming to identify and map various biological entities, including protein types, disease types and cell types.

There are several reasons for developing different metadata standards: complexity, diversity, increasing data generation and inter-relatedness (Barrett et al., 2012), all of which can vary



FIGURE 2.4: Terminologies for various metadata types (Mayernik, 2021)

depending on the repository's goal. The BioProject and BioSample databases (Barrett et al., 2012) were developed to capture metadata for different types of biological data stored in databases such as ArrayExpress and GEO. The EBI BioSamples metadata includes details such as data type (e.g., genome sequencing), attributes, methods, project aim, title and a brief description.

A widely used biomedical dataset discovery system, DataMed, uses a specific metadata ingestion pipeline designed to extract, index, and map data using the DATS (Chen et al., 2018). The DATS metadata standard aims to extract generic and applicable elements for any type of biomedical dataset, including properties that describe entities, such as information about the materials used. These include metadata information like publication, software, data repository and access method. Figure 2.5 presents an example of metadata as displayed in DataMed.

The absence of metadata standardisation (Hughes et al., 2023) and structured metadata (Johns et al., 2023) that describes the history of a dataset's construction (provenance metadata) can lead to issues with reproducibility. Several studies confirm the importance of provenance metadata in reproducibility and trustworthiness tasks (Gierend et al., 2023; Samuel and König-Ries, 2022; Gierend et al., 2024). Therefore, the growing demand for this type of metadata has advanced the development of several provenance models.

The screenshot shows the DataMed website interface. At the top is a dark blue header with the 'DataMed' logo on the left and navigation links (Home, About, Repositories, Search, Web API, Feedback) on the right. Below the header is a light green bar with a '< Go Back' link. The main content area is divided into two sections: 'Metadata' and 'Distributions'. The 'Metadata' section contains the following information:

- Name:** Second Primary Cancers in Patients With Castration Resistant Prostate Cancer
- Repository:** ClinicalTrials.gov
- Identifier:** clinicaltrials/NCT02702908
- Description:** This study aims at estimating the incidence of second primary malignancies as well as the overall survival among metastasized prostate cancer (mPC) and metastasized castrate-resistant prostate cancer (mCRPC) patients not treated with radium-223-dichloride.
- Data or Study Types:** clinical trial
- Source Organization:** Unknown
- Access Conditions:** available
- Year:** 2016
- Access Hyperlink:** <https://clinicaltrials.gov/ct2/show/NCT02702908>

The 'Distributions' section contains a single bullet point: Encoding Format: HTML ; URL: <https://clinicaltrials.gov/ct2/show/results/NCT02702908>

On the right side of the 'Metadata' section, there is a word cloud with the following words: estimating, aims, cancer, metastasized, primary, prostate, second, study, mpc, well.

FIGURE 2.5: An example of DataMed metadata includes several pieces of information about the dataset, such as its name, repository, identifier, and more.

2.3 Provenance

One critical type of metadata is provenance information (Hartig and Zhao, 2010). While the definitions of provenance differ based on the domain, it generally defines all steps involved in building a dataset (Johns et al., 2023). Data provenance is a type of metadata that explains the origin, history and changes of data (Simmhan et al., 2005). A simple understanding of the concept of “provenance” refers to explaining what happened to the data (Curcin, 2017). Johns et al. (2023, p.2) defines provenance as “the origin, processing, and movement of data”. Several studies emphasise that provenance information is strongly associated with the interpretation and reproducibility of results (Liu et al., 2020; Gierend et al., 2024; Johns et al., 2023; Gierend et al., 2023).

Several efforts have been made to understand and establish the foundations of provenance. Herschel et al. (2017) presented a classification of provenance in a hierarchy of four types, as illustrated in Figure 2.6. The first type is provenance metadata that encompasses information about

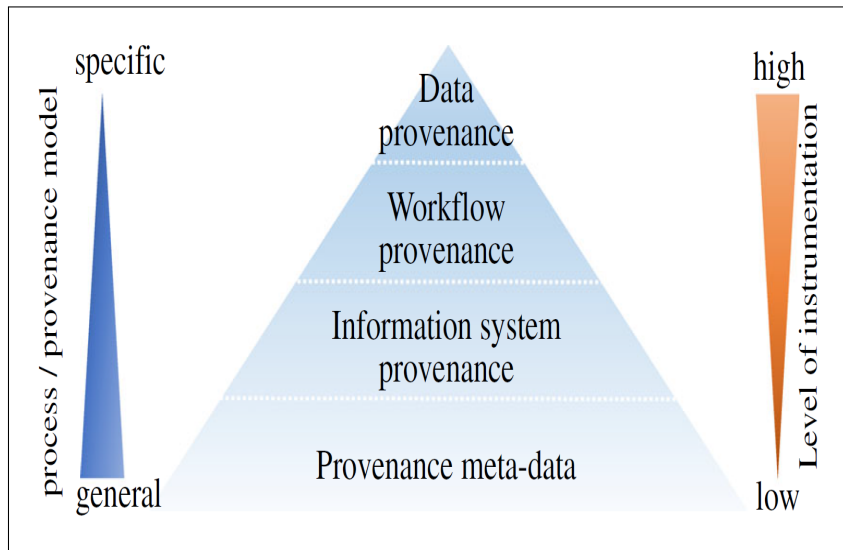


FIGURE 2.6: Hierarchy of provenance metadata types (Herschel et al., 2017)

any production process. Users have the broadest freedom to model, store, and access provenance information for any product type. The second type is information system provenance, which refers to metadata describing processes in information systems involved in the dissemination of information, such as communication (Deutch et al., 2014; Macko and Chiarini, 2011), storage and retrieval (Chirigati et al., 2016) or distribution (Ko and Will, 2014; Zhou et al., 2012). The provenance collection in this context includes the inputs, outputs, and parameters of each process. The third type is workflow provenance, where a workflow is considered in Herschel’s survey (Herschel et al., 2017) as a graph, with nodes representing functions or modules and edges representing a predefined data flow between the modules. The final type is data provenance, which involves tracking the processing of single data components with the highest resolution. This means capturing provenance at the level of data items and the operations they undergo.

Provenance has been utilised for decades in computer science, with the aim of tracking the origin of data and the processing steps involved in generating final results to verify their quality (Goble, 2002). From a data management perspective, data management systems focus on identifying various types of information, including the source of the data within databases (where provenance), the justification for the presence of the data (why provenance) and the data relevant to observed outcomes (how provenance) (Green et al., 2007; Cheney et al., 2009). An overview by Gierend et al. (2024) highlights that the concept of provenance is essential in various scientific research fields, including environmental studies (Liu et al., 2020), biomedical and health research domains, material sciences, and security and privacy (Pan et al., 2023). As confirmed by Ahmed et al. (2023), data provenance has been utilised in various fields, including artificial intelligence, e-services and healthcare.

Data provenance is used for several tasks, including ensuring reliability (Baum et al., 2017; Groth and Moreau, 2013), quality (Groth and Moreau, 2013), reproducibility (Liu et al., 2020; Gierend et al., 2023), trustworthiness (Gierend et al., 2023; Groth and Moreau, 2013), enhanced interpretability (Gierend et al., 2023) and data tracking (Simmhan et al., 2005; Johns et al., 2023). Additionally, Herschel et al. (2017) note that provenance is applied in various use cases, such as supply chains, complex data processing and scientific experiments. Furthermore, data provenance addresses the lack of documentation related to data processing and experimental parameters (Johns et al., 2023). Provenance can assist in identifying the origin of data as well as the processes performed on it over different phases and periods of time (Ahmed et al., 2023).

2.3.1 Provenance Models

To represent and store provenance information, several models have been introduced across scientific and computational domains. The provenance research community has begun to explore provenance capabilities and how they can be represented across different systems (Samuel, 2019). These models provide structured vocabularies and data structures to describe how entities (such as data or files), activities (such as processes or computations) and agents (such as users or systems) interact.

One of the earliest formal models developed for this purpose was the Open Provenance Model (OPM) (Moreau et al., 2011). This model consists of three main specifications: nodes, dependencies and roles. First, a node relies on three fundamental concepts: artefacts, processes and agents, as illustrated in Figure 2.7. An artefact refers to a piece of state; a process represents an activity or a set of activities; and an agent is the entity that acts as a catalyst for the process. Second, dependencies in this model express how the core entities are related. Finally, roles describe the role of each entity involved in the process.

Following this, the W3C Provenance Incubator Group was established to develop and understand the requirements for provenance information across different scientific domains. Additionally, this community aimed to standardise the provenance model. Subsequently, the Provenance Working Group developed a recommendation known as PROV, comprising a set of W3C recommendations. This set includes eight specifications for exchanging provenance information between heterogeneous environments (Groth and Moreau, 2013). Moreau et al. (2015) discussed the main requirements and principles for designing PROV, which is built around three core concepts, as illustrated in Figure 2.8: Entity, representing conceptual, digital, physical, or any thing in the world; Activity, describing how entities emerge and change over time; and Agent, responsible for performing activities or managing entities.

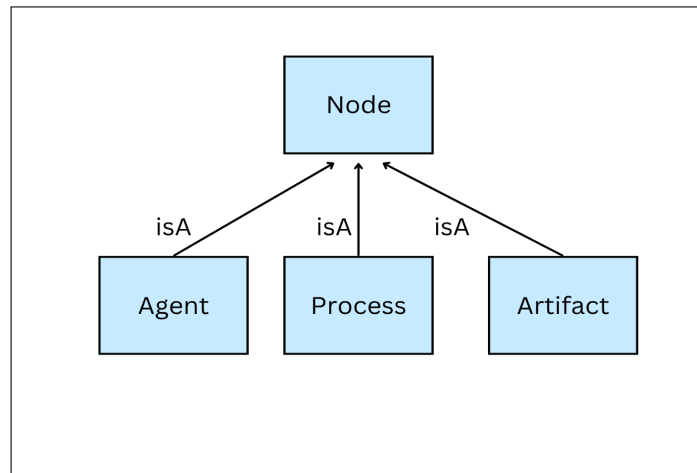


FIGURE 2.7: The core concepts of OPM (Moreau et al., 2011)

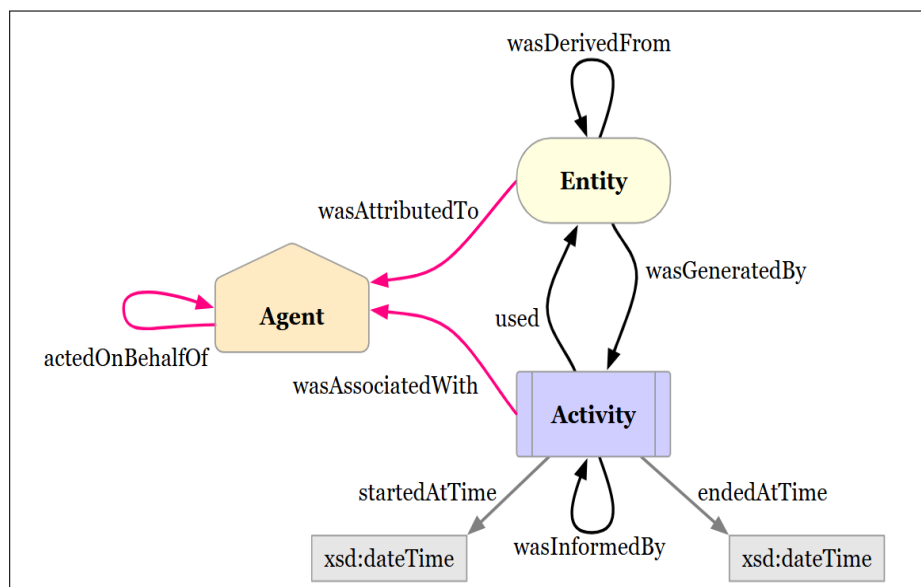


FIGURE 2.8: The core concepts of PROV: Entity, Activity and Agent (Moreau et al., 2011)

Other provenance models were adapted and developed for specific communities and domains. In addition, several provenance models were extended from OPM and PROV, primarily focusing on scientific workflows. For instance, P-Plan (Garijo Verdejo and Gil, 2012) was extended from the PROV model to represent plans and their prior executions of abstract scientific workflows. Another example that extends PROV, OPM and P-Plan is OPMW (Garijo and Gil, 2011). It is used to capture two types of provenance information: prospective and retrospective, by connecting instances, templates, and workflow executions. However, OPMW causes an overload of OPM concepts without introducing added vocabularies (Missier et al., 2013b); thus, they developed another model, D-PROV, to track the provenance traces in workflows.

Although several provenance models exist, Zhao et al. (2012) found that workflow decay may still occur, even in systems designed to capture provenance information. In their experiment

using Taverna workflows — a workflow system that captures provenance — within the public repository myExperiment, they observed failures in both workflow execution and in reproducing consistent results. [Baum et al. \(2017\)](#) confirms that several provenance issues remain to be addressed in life sciences, even with the presence of the W3C framework.

In this thesis, we conducted a study to review existing data provenance models, including both generic and domain-specific approaches. This study was undertaken as a brief systematic review, following three main stages: planning the review, conducting the review, and reporting the results. The process followed the guidelines proposed by ([Kitchenham et al., 2004](#)).

The first stage, planning the review, aimed to identify the review’s requirements and to establish a structured process. The primary objective was to identify existing provenance models. A review protocol was developed to ensure an accurate and reliable methodology, including a search strategy and inclusion and exclusion criteria. The inclusion criteria required studies to present and describe general provenance models that focus on, or are relevant to, the biomedical domain. We excluded studies that (1) do not focus on provenance models, (2) do not clearly explain the models or (3) are published in a language other than English.

In the second stage, we conducted the review in accordance with our established protocol. The review process was carried out between February 2023 and June 2023. We focused on surveys, review papers and publications from the International Provenance and Annotation Workshop (IPAW) and the annual Theory and Practice of Provenance (TaPP) workshop. As a result of the search and selection steps, we initially identified 61 papers related to provenance models. After applying our inclusion and exclusion criteria, a total of 18 papers remained, all of which described provenance models focused on, or related to, the biomedical domain.

In Table 2.1, we summarise each model or extension by including the following details: (1) an identifier for each model; (2) the model name; (3) the target research community; (4) the data that the model focuses on; (5) the data type (e.g., microarray data); (6) a brief description of the model.

2.3.2 Provenance in Biomedicine

In the field of biomedicine, biomedical data are typically collected in various formats and types for different purposes, one of which is to advance healthcare and research ([Johns et al., 2023](#)). This data may include information on treatments, conditions, and experimental outcomes, providing either certain measurements or abstract observations. The origin (provenance) of such data should be captured accurately to preserve its meaning and reliability.

Although collecting data provenance is popular in several communities, such as geoscience and computer science, this issue still persists in other data-driven research domains, including biomedical research (Buneman and Tan, 2007; Collins and Tabak, 2014; Baum et al., 2017; Liu et al., 2020; Johns et al., 2023; Gierend et al., 2024). Only a few contributions have emerged in developing provenance data models for biomedical domains to address this gap.

McCusker and McGuinness (2010b) developed a provenance extraction technique. This work aimed to transform MAGE (MicroArray and Gene Expression) metadata, used for representing microarray experiments, into the OPM and PML models. The approach involved converting experimental data in MAGE-TAB format into RDF, represented within the OPM and PML frameworks. However, it focuses solely on capturing the data provenance of microarray experiments. Additionally, it was validated using a small dataset, raising concerns about its scalability to all high-throughput biomedical data.

Another approach, the ISO 23494-2 Common Provenance Model, aims to standardise provenance information for biological specimens and data (Wittner et al., 2021). The key objectives of this model are to evaluate the quality of such data, facilitate its reproducibility and ensure its integrity. The model involves six steps: designing concepts and requirements, structuring on the W3C PROV model, identifying the needs for provenance information related to biological material or specimen acquisition, generating provenance data, identifying the provenance of computational workflows and addressing the security extensions of provenance. The design of this model is tailored for specialists in HW/SW systems used in the domains of biotechnology and biomedicine. Wittner et al. (2024a) confirm that the development of the standard is still ongoing, in enquiry phase with ISO members and invite experts and researchers in biotechnology and biomedicine to contribute to its advancement.

ProvCaRe was developed as a framework specifically designed for clinical and healthcare research by (Sahoo et al., 2019). The architecture of the ProvCaRe system is presented in Figure 2.9. The framework is primarily concerned with extracting provenance information associated with published studies in the field of sleep medicine. This system relies on the ProvCaRe S3 model, which includes semantic provenance metadata and a knowledge base derived from the National Sleep Research Resource (NSRR). This model is based on two well-known clinical frameworks: Population, Intervention, Comparison and Outcome (PICO) model and Ontology for Clinical Research (OCRe) ontology. The PROV-O ontology was utilised during the implementation phase of this model. The aim of the S3 model is to identify three key components within sleep medicine publications: study methods, study tools and study data. The first component, study methods, includes various types of information about the study design, such as whether it is interventional or observational, as well as details on recording techniques and statistical analysis

methods. The second component consists of the tools used in the study, including the instruments employed during data recording and analysis. The final component represents the study data, such as information about the variables.

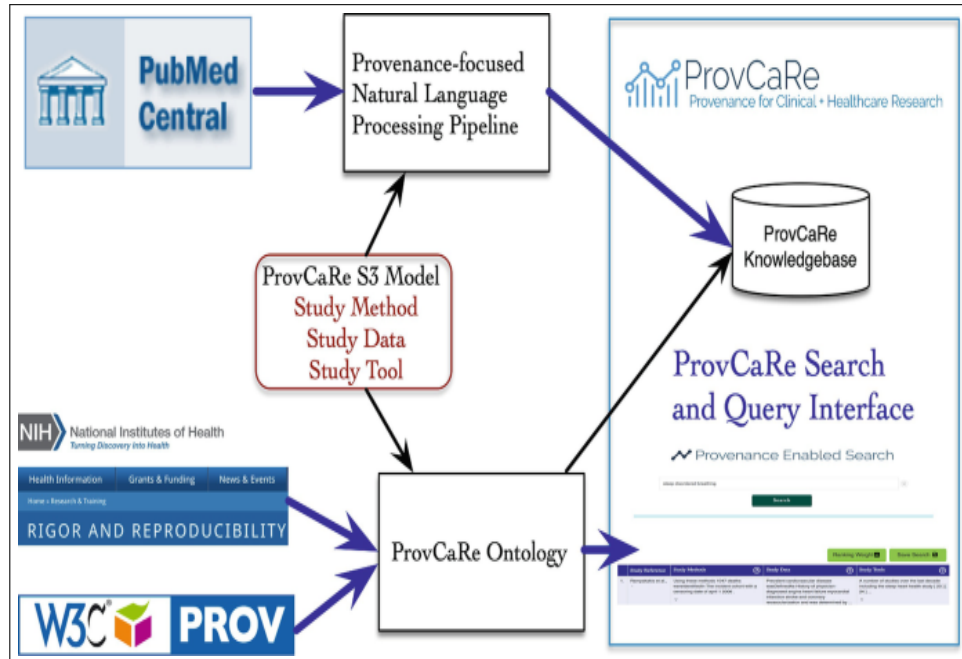


FIGURE 2.9: Provenance S3 model (Sahoo et al., 2019)

Later, the same team behind the Provenance project developed an ontology model that integrates neuroimaging data with the S3 model, referred to as NeuroBridge (Sahoo et al., 2023). This aims to extract provenance information from neuroimaging data experiments. The team used the same framework as the Provenance project, restructuring and expanding the Provenance ontology. However, this model is unable to cover certain neuroimaging data, including spinal imaging studies and brain tumor scans. It also fails to capture essential information, such as quality assurance steps and statistical results (Sahoo et al., 2023).

While the efforts to develop provenance models in biomedicine are commendable, these models, extensions and frameworks may not fully meet the provenance requirements across all biomedical fields. It could be due to various types of data, scalability and various used techniques. In addition, the generality of some provenance models poses a challenge. As observed by Sahoo et al. (2019), existing approaches in the clinical and health domains provide only limited support for capturing provenance information in accordance with PROV specifications.

Leipzig et al. (2021) highlighted that the Provenance project (Valdez et al., 2017) cannot completely address the gap between human-readable protocols and machine-readable metadata formats. This project is now inoperable and is specifically modeled for the sleep medicine domain.

It has been observed that NeuroBridge (Sahoo et al., 2023) faces scalability challenges due to its reliance on the time-intensive process of ontology engineering. It also fails to capture essential information that is needed in reproducing neuroimaging experiments.

As a result of these obstacles, such provenance models and tools may not be widely adopted in biomedical domains. As Gierend et al. (2024) confirmed in their scoping review, there is a heterogeneity of methods and models with varying feature sets. This can provide several opportunities to address these gaps effectively.

2.3.3 Workflow Visualisation

While formal provenance models such as PROV-DM provide serialisations that can be used to capture provenance components, some scientific domains particularly those focused on scientific workflows often employ alternative forms of visual representation to convey research procedures and data processing steps. These workflow visualisations typically integrate both manual and computational steps into structured diagrams that support understanding, reproducibility and reusability (Willoughby and Frey, 2017).

In the domain of bioinformatics and biomedical data analysis, visual workflow systems have played a central role in managing data pipelines (Wratten et al., 2021). According to Di Tommaso (2017), several existing tools, such as Galaxy (Goecks et al., 2010) and Taverna (Oinn et al., 2004), provide visual interfaces to execute and share workflows, thereby supporting reusability. The purpose of these visual systems is to help users track dependencies between tools and datasets, while also facilitating documentation and provenance capture during pipeline execution.

The use of visual representations for workflows has become increasingly common in various domains, including business process modelling. Von Rosing et al. (2015) present Business Process Model and Notation (BPMN), a standard that describes business process semantics using graphical notation. Its objective is to support business process modelling for both technical and non-technical users, enabling the representation of complex processes. BPMN offers a high level of expressiveness and can also be used for related tasks, such as specifying business rules (Milanovic et al., 2008) and business events (Decker et al., 2007). Other examples of visual process modelling languages include flowcharts, UML activity diagrams, Petri nets and event process chains (Kocbek et al., 2015).

In scientific communication, Graphical Abstracts (GA) have emerged as a method of visually summarising a study's design, methodology or key findings. GAs are often included in scholarly articles to facilitate browsing and assist readers in selecting relevant papers (Yoon and Chung,

2017). According to Lazard and Atkinson (2017); Bredbenner and Simon (2022), the use of GAs is required by some journals to convey the essence of research. For example, GAs are frequently employed in medical journals and by professional organisations (Millar and Lim, 2022). However, this technique is considered less formal than workflow languages or provenance ontologies (Millar and Lim, 2022).

2.3.4 Provenance for Biomedical Dataset Search

As mentioned earlier, there is a significant need for capturing provenance in biomedical research, as it assists researchers in finding, accessing, interoperating, and reusing datasets. In addition, data provenance is a type of metadata that is crucial in biomedical domains, as it explains the history of the provided data, which serves several purposes, including reproducibility. Moreover, metadata is important in dataset search, as it assists researchers in selecting relevant datasets based on their needs in biomedical domains. However, the provision of data provenance for existing datasets remains a gap that needs to be addressed, as it may support biomedical researchers in searching for and selecting appropriate datasets. Figure 2.10 presents the intersections between these three main domains and illustrates the gap.

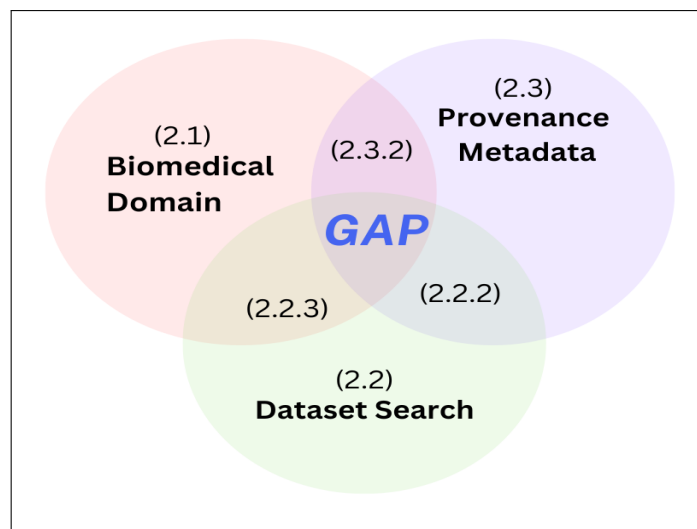


FIGURE 2.10: Gap at the intersection of the three main domains.

2.4 Improving Provenance Annotation

Several approaches have been proposed to capture and collect provenance. In addition, various studies have been conducted to improve and categorise these approaches. Blount et al. (2021) classified the capture and collection of provenance metadata into two key categories: observed

	Model	Community	Targeted Data	Data Type	Description
M1	W3C PROV (Missier et al., 2013a)	General	Provenance data on the Web	The origin or source of information	The PROV family includes various specifications, such as PROV-O and PROV-N. Its core is PROV-DM, which aims to capture the essential information of provenance.
M2	W7 Model (Ram et al., 2009)	General	Provenance data	The semantics of data provenance	This model focuses on capturing seven elements, which are "what," "when," "where," "how," "who," "which," and "why."
M3	MIAME (Brazma et al., 2001)	Life Sciences	Experimental data	Microarray data	The objective of this model is to capture the provenance associated with microarray-based gene expression experiments. It collects various elements, such as the experimental design, array configuration and sample details.
M4	Common Provenance Model (CPM) (Séroussi et al., 2022)	Life Sciences	Process data	Experiments	It aims to collect provenance information based on the PROV and provenance composition patterns.
M5	Biotechnology (Wittner et al., 2021)	Biology and Biomedical Researchers	Workflow data	Biological material and data	The aim of this model is to collect and standardise provenance information based on ISO and PROV standards. It captures several elements of provenance for various purposes, including quality assessment, research reproducibility, and error tracking.
M6	High Throughput Experiments and Provenance (McCusker and McGuinness, 2010a)	Biomedical informatics	Experimental workflow data	Microarray experiments	The aim of this model is to collect provenance information based on two models: MAGE and OPM.
M7	LONI Pipeline Provenance (MacKenzie-Graham et al., 2008)	Biological science	Workflow data	Neuroimaging	This model captures the provenance of neuroimaging workflows, collecting several elements such as the name, subject ID, and birth and death dates.
M8	A Comprehensive Model for Provenance (Sultana and Bertino, 2012)	System developers	Any data object	Abstraction layer of workflow/process/OS	The goal of this model is to capture provenance information at two levels: the application level and the system level.
M9	ProvONE (Missier et al., 2015)	DataONE (scientific community)	Earth observational data	Data Observation Network for Earth	The goal of this model is to collect provenance information based on a proper extension of the PROV model.
M10	SC-PROV (Markovic et al., 2015)	General	Workflow data	Social computations	It aims to capture the provenance of social computations. The objective of this model is to capture the provenance associated with social computations. It collects various elements of workflow steps, such as activities, based on extensions of PROV-O and P-PLAN.
M11	CRISTAL (McClatchey et al., 2015)	Medical researchers (Neuroscientists)	Workflow data	Research analysis processes for Alzheimer's disease	The aim of this model is to capture the provenance of the neuGRID and N4U projects. It focuses on collecting provenance information from both the specification and execution stages of the analysis workflows.
M12	Prov Viewer (Kohwalter et al., 2016)	General	Scientific workflow data	Scientific data	Its goal is to capture provenance data in order to generate a provenance graph. It focuses on collecting activities, entities, and agents.
M13	Neuroscience Experiments Systems (NES) (Ruiz-Olazar et al., 2016)	Neuroscience	Experimental data	Experimental raw data of Neuroscience	This model aims to capture provenance information related to experimental raw data, using the 7Ws model as a framework.
M14	DfA-prov (Silva et al., 2018)	Computational Scientists	Input data and parameters of the CSE function	Data of high-performance applications	The aim of this model is to collect the provenance of databases during or after the implementation of a CSE application, based on the PROV-DM standard.
M15	Semantic Sensor Network System Deployment Provenance Ontology (SDPO) (Silva et al., 2018)	Internet of Things (IoT)	Describing deployment data	Deployment data of IoT systems	The aim of its model is to capture the provenance of IoT systems during its developments, based on PROV.
M16	Provenance Data Model for Astronomy (Galkin et al., 2018)	Astronomy	Astronomical dataset	Astronomical data	The aim of this model is to capture the provenance of astronomical datasets. It is based on two frameworks: IVOA and PROV.
M17	OpenSoils (da Cruz et al., 2018)	Agriculture	Experimental data	Soil experiments	The aim of this model is to capture the provenance of experiments at various levels or layers. It is based on three frameworks: FAIR, PROV, and semantic web approaches.
M18	ProvCaRe (Liu et al., 2020)	Clinical and Healthcare Research and biomedical research	Experimental data	Sleep medicine research studies	It aims to extract, capture, and analyse provenance information from publications, based on the S3 model, which includes study methods, study tools, and study data.

TABLE 2.1: Existing provenance models and extensions

provenance and possible provenance, and we expand upon other techniques for this tasks. Observed provenance is captured when observable changes occur in system events, whereas possible provenance refers to provenance generated without the restrictions imposed by observable actions within the system. In this study, we focus on the latter category. Table 2.2 presents all studies conducted under each approach within the updated dendrite.

Uncertain approaches to provenance incorporates a measure of uncertainty and assumes that events occur with a certain probability. Few studies focus on this approach. [Huang and Fox \(2004\)](#) proposed a technique to explore uncertainty in events that cannot be measured. It focuses on both the uncertainty associated with trust in relationships and facts. This work was later extended to represent provenance using the W3C PROV format ([De Nies et al., 2013a](#)).

[Blount et al. \(2021, p.3\)](#) defines inferred provenance as “the reconstruction of provenance metadata, which is inferred based on user behaviour or underlying system rules”. [Magliacane \(2012\)](#) developed a methodology for reconstructing provenance, aiming to automate the process, reconstructing provenance and using file contents. This approach was based on machine learning techniques, including the deep reasoning approach. The goal was to evaluate the provenance information of several documents in a shared folder and demonstrate it through a small pilot. Another similar approach to reasoning provenance inference was proposed by [Kodagoda et al. \(2017\)](#). This work employs machine learning instead of relying on logs of low-level interaction.

A few contributions have emerged in inferring provenance from publications. [Sahoo et al. \(2019\)](#) proposed a framework specifically for clinical and healthcare research, called ProvCaRe, which is detailed in Subsection 2.3.2. It is based on a Named Entity Recognition (NER) module, which incorporates various techniques, including a knowledge model that utilises OWLAPI, the MetaMap tool and the Open Biomedical Annotator tool.

[Sahoo et al. \(2023\)](#) developed NeuroBridge, designed to infer provenance information from neuroimaging data experiments. It employs an ontology model that integrates neuroimaging data with a custom-developed data model. However, this model is unable to cover certain neuroimaging data, including spinal imaging studies and brain tumour scans.

Although there have been several contributions to information extraction using LLMs ([Gutierrez et al., 2022](#); [Cao et al., 2023](#); [Dagdelen et al., 2024](#)), only one study specifically aims to infer the provenance of data preparation pipelines ([Lauro et al., 2024](#)). This contribution, PROLIT, seeks to automatically track and capture granular provenance information related to the steps of data transformation within input pipelines written in Python. This work seeks to capture provenance information from scripts and workflows, but it is unable to capture or leverage provenance from publications.

Provenance	Observed	In-situ	Automatic (Pasquier et al., 2017; Murta et al., 2015; Sáenz-Adán et al., 2018; Pina et al., 2024; Chapman et al., 2022)
			Annotated (Lerner and Boose, 2014; Guedes et al., 2018; Interlandi et al., 2015; Parulian and Ludäscher, 2023)
		Post-hoc	Replay (Stamatogiannakis et al., 2017; Thurler et al., 2025)
			Logs (De Nies et al., 2013b; Pasquier et al., 2018; Zengy et al., 2022; Cheng et al., 2024)
	Contextual (McPhillips et al., 2015; Zhang et al., 2017)		
	Possible	Uncertain	Uncertain (Huang and Fox, 2004; Idika et al., 2013; De Nies et al., 2013a)
		Inferred	Abductive reasoning (Kodagoda et al., 2017)
			Machine learning (Magliacane, 2012; Kodagoda et al., 2017; Pocock, 2021)
			NLP (Sahoo et al., 2019, 2023)
			LLM [our approach], (Lauro et al., 2024)

TABLE 2.2: Overview of studies under each approach represented in the provenance dendrite diagram.

2.4.1 Improving Other Metadata Annotations

Several efforts have emerged to improve metadata annotation across various scientific domains, including biomedical domains. [Leo et al. \(2024\)](#) developed an approach called Workflow Run RO-Crate, which aims to capture provenance information related to the execution of computational workflows. This model is an extension of RO-Crate (Research Object Crate) ([Soiland-Reyes et al., 2022](#)), designed to provide research artefacts, such as data, software and workflow results, with metadata that describes their context and history. The objective of Workflow Run RO-Crate is to capture provenance at multiple levels of granularity. Although its profiles are still evolving, the approach has already been implemented in several workflow systems, including Galaxy, CWLProv, Snakemake and Nextflow, demonstrating its growing maturity and practical relevance.

[Wittner et al. \(2024b\)](#) describe a provenance implementation developed as part of the BY-COVID project, which applies existing standards — including ISO 23494 and Workflow Run RO-Crate — to support the structured capture of provenance information in the context of biomedical data

analysis. The implementation focuses on capturing provenance across organisational boundaries, for example: samples collected in a hospital, data generated from these samples in a laboratory and subsequent data processing conducted by individual researchers or research groups. It also includes provenance information relating to the precursors of digital objects, such as datasets. However, although the deliverable presents a concrete use case within BY-COVID, its scope was limited to specific aspects of data processing, integration and analysis, and did not extend to the detailed capture of provenance associated with sample acquisition in clinical settings.

[Gil et al. \(2015\)](#) developed a project called OntoSoft, aimed at capturing metadata about scientific software, specifically geoscience software. The system enables geoscience researchers to obtain more detailed information about existing software to support its reuse. This work assists in recording several basic metadata elements about software, including its identification, function, execution, application in research and other related information. OntoSoft is based on a dedicated ontology—the OntoSoft ontology—which was developed to capture and structure this metadata. However, this work does not enable the capture of metadata in a machine-readable manner ([Garijo et al., 2019](#)).

Investigation/Study/Assay (ISA) is a software tool developed to annotate experimental metadata from high-throughput studies ([Rocca-Serra et al., 2010](#)). It is based on the ISA-Tab format, which was designed to structure and communicate metadata effectively ([Sansone et al., 2008](#)). The ISA framework captures three main components: Investigation, which includes information to understand the overall aim of the experiment; Study, which involves information about the subject used in the study; and Assay, which relates to the study and consists of measurements and the technologies employed.

2.5 Large Language Models (LLMs)

In recent years, remarkable advancements have been achieved in language models, mainly attributed to the development of several techniques, including transformers ([Chernyavskiy et al., 2021](#)), enhanced computational capabilities and the wealth of large-scale training data ([Naveed et al., 2023](#)). Consequently, significant progress in LLMs has led to the development of AI systems capable of processing and generating texts ([Naveed et al., 2023](#)). In addition, LLMs demonstrate immense promise in disciplines that integrate human expertise with AI techniques.

LLMs are known as modelling methods used to predict the next tokens in a sequence, which demonstrated their success in integrating knowledge into a model (Zhao et al., 2023). Taveekitworachai and Thawonmas (2023) defines LLMs as language models with an appropriate architecture trained on large amounts of data. Those models depend directly on NLP techniques as well as other types of ML techniques.

Earlier models, such as mT5 and T5, relied solely on transfer learning techniques (Raffel et al., 2020; Xue, 2020). However, with the advent of GPT-3, LLMs demonstrated the ability to deal with zero-shot transferable tasks and downstream tasks without the need for fine-tuning techniques (Naveed et al., 2023). Later, these models have improved to perform tasks more effectively in zero-shot settings than in few-shot settings by using fine-tuning with task instructions (Sanh et al., 2021).

These LLMs can perform various tasks that approximate human-level performance (Wang et al., 2019a). Subsequently, with advancements in this domain, LLMs have demonstrated emergent capabilities, including reasoning, decision-making, and in-context learning (Naveed et al., 2023). Many studies have proven that LLMs have a significant capability to effectively perform a wide range of activities that meet human expectations such as LLMs can be used in question-answering (Gilson et al., 2022; Kasneci et al., 2023; Taveekitworachai and Thawonmas, 2023), summarising texts (Nordgren and E Svensson, 2023), generating code snippets (White et al., 2023), generating texts (Chen et al., 2023) decision-making process (Liu et al., 2023) and other language-related tasks.

Recent advancements in LLMs have led to the development of a wide range of both general-purpose and domain-specific models. Table 2.3 provides a summary of several existing LLMs.

Model	Domain	Description	Strengths	Limitations
GPT-4o (Hurst et al., 2024)	General	A multimodal model with strong capabilities in zero-shot and few-shot prompting.	Strong performance across many tasks	General-purpose, not fine-tuned on scientific texts
SciBERT (Beltagy et al., 2019)	Scientific	Trained on scientific text from Semantic Scholar.	Domain relevance	Smaller size; task-specific fine-tuning often required
BioBERT (Lee et al., 2020)	Biomedical	Pretrained on PubMed and PMC articles.	Biomedical relevance	Less generalisable beyond biomedical text; trained on only PubMed abstracts
PubMedBERT (Gu et al., 2021)	Biomedical	Trained entirely on PubMed abstracts.	Biomedical specificity	Narrow domain scope; trained on only PubMed abstracts
LLaMA2 (Touvron et al., 2023)	General	Open-source models with competitive performance.	Transparency, reproducibility	Require fine-tuning for domain use

TABLE 2.3: Comparison of LLMs

2.5.1 LLMs in Biomedicine

Extracting data from biomedical texts requires genuine human effort and can be time-consuming. Consequently, several approaches in biomedical text mining have been developed to address the challenge of extracting information from unstructured literature, including named entity recognition (NER) (Zweigenbaum et al., 2007) and the extraction of metabolic reactions (Czarnecki et al., 2012). While the observed developments between technology and biomedicine have revolutionised the biomedical text mining domain, several obstacles have impeded the development of text generation and mining. These obstacles include the complexity of the field's data and the vast amount of existing data being generated (Chen et al., 2023).

Several efforts have been explored for utilising LLMs in biomedical text generation tasks. Numerous studies (Chen et al., 2023; Thapa and Adhikari, 2023; Tian et al., 2024) confirm that LLMs generally demonstrate impressive performance across various biomedical domains and tasks. Several domain-specific models have been developed and trained on biomedical datasets, including BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019). Additionally, other biomedical models have employed mixed-domain training and domain-specific training techniques, such as PMC-LLaMA (Wu et al., 2024).

While these models have proven their effectiveness in biomedical domains, their efficiency might be affected by limitations such as adaptability to biomedical languages. Sahoo et al. (2024) highlights that such these techniques face a lack of depth in several specific areas, which necessitates extensive resources for training. Fedus et al. (2022) confirm that training such models can be performed either from scratch or from a checkpoint, both of which are resource-intensive as the model size increases significantly. Additionally, there is a risk of overfitting such models, resulting in poor generalisation (Sahoo et al., 2024).

One of the latest surveys (Sahoo et al., 2024) on the use of LLMs in biomedicine highlights a positive consensus regarding their performance when trained with improved data quality and enhanced prompting methods (Wei et al., 2022a; Chowdhery et al., 2023; Touvron et al., 2023). However, Sahoo et al. (2024) highlight several limitations in using LLMs for biomedical NLP applications. First, high-quality data is required, as it has a significant effect on the performance of LLMs. Second, there is an emphasis on the importance of employing fair approaches to evaluate the performance of LLMs, which remains an emerging challenge across biomedical NLP tasks. Third, there is a need to improve the transparency and quality of results by explaining the steps involved in generating them.

In this thesis, we selected GPT-4o for several reasons. GPT-4o does not require additional training or fine-tuning, as it has been pretrained on a vast corpus of internet-scale data and is optimised for strong zero-shot and few-shot performance (OpenAI, 2024), as detailed in Section 2.6. While models such as SciBERT and BioBERT offer advantages within specific domains, they require fine-tuning and lack the interactive capabilities of more general models. Moreover, recent evaluations (Chang et al., 2024) have shown that general-purpose LLMs, including GPT-4 and GPT-4o, outperform domain-specific models on many complex reasoning tasks, even in biomedical contexts. We further explain the use of LLMs in NLP and their relevance to this research in Section 3.4.1.

2.5.2 Summary

Despite the significant use of LLMs in recent years, they have several limitations and errors. Those limitations and errors can be due to the absence of innate comprehension of the world and their reliance on patterns acquired from data (Nordgren and E Svensson, 2023). For instance, LLMs can generate plausible language but are sometimes inaccurate or nonsensical. As a result of such limitations and errors in LLMs, some studies mention their trustworthiness is low (Si et al., 2022), and they have difficulties when dealing with complex tasks (Bender et al., 2021). Moreover, LLMs can perpetuate biases because they are trained on data publicly available on the internet (Nordgren and E Svensson, 2023). In order to overcome some of those issues, we might use prompt engineering techniques or consult with experts in LLMs. Various studies have emphasised that task-specific training or enhanced prompt techniques should be used to improve the performance of LLMs (Parisi et al., 2022; Giray, 2023; Zhang et al., 2023a).

2.6 Prompt engineering

Prompt engineering is defined as a set of methods and techniques to design, process and refine a collection of instructions or questions (Ekin, 2023). The main goal of using prompt engineering in LLMs is to optimise and elicit more precise, accurate, replicable responses for various levels of tasks (Wang and Jin, 2023). A well-designed prompt, as mentioned by (Wang and Jin, 2023; Ekin, 2023), is crucial skill for user interaction with LLMs, leading to more precise and relevant responses and improved outcomes. Several studies emphasise the effectiveness of using prompts in retrieving information from LLMs (Schick and Schütze, 2021; Schmidt et al., 2024).

Tan et al. (2023) refers to the importance of prompt engineering when implementing complex tasks, which includes several steps/ sub-tasks and long text descriptions. Therefore, several

recommendations were provided for better LLM prompts by Meskó (2023) that can enhance the interactions with LLMs: being more specific; using different prompts for same task; identifying the kind of outputs; iterating and refining the prompts; utilising open-ended questions; using prompting methods based on the task; considering temporal awareness if required in the task; describing settings and providing context around the task. Recent studies have further identified variations in performance between simple and complex prompts when employing LLMs for NLP tasks (Liu et al., 2023; Peng et al., 2023; Jiao et al., 2023).

Several recent prompt strategies were established in order to improve the prompt used in LLMs. In the following section, we highlight the most prominent strategies proposed:

2.6.1 Zero-Shot Prompt

This strategy aims to provide an initial description of a task in the prompt for LLMs to get started with. The description could be a single and comprehensive prompt without any further information or context (Kojima et al., 2022). Based on this prompt, a response would be generated. This prompt strategy is typically utilised when LLM training is unnecessary. For example, this strategy can be used for translating or summarising texts, which can be performed with the internal knowledge of the LLM.

Despite the fact that LLMs have shown a powerful result when using zero-shot prompts, they still face obstacles when performing complex tasks. Therefore, **the Few-Shot prompt** can be used to overcome those limitations because of their technique, which enables in-context learning (Shen et al., 2023).

2.6.2 Few-Shot prompt

By providing more explanations in the prompt, the LLM can be guided towards a better result by conditioning it with examples. In other words, few-shot prompts include more detailed information or more examples about the task. It is designed to provide more than one example corresponding to a task. Brown et al. (2020) suggests that providing several examples can assist in interpreting the details of the task.

An example of a complex task is sentiment classification, which requires providing a few examples to generate a more accurate result. Nevertheless, it might not be appropriate for more complex reasoning problems because those problems need to perform several reasoning steps

(Shen et al., 2023). This problem can be solved by breaking down the reasoning problem into several steps.

2.6.3 Chain-of-thought (CoT) prompt

This strategy was designed to break reasoning several steps, which is a remarkable technique to be used in LLMs. It aims to address the limitations of few-shot prompt. However, this strategy still faces obstacles while engaging in complex reasoning or when the tasks are long but there is not enough explanations or examples (Wei et al., 2022b).

Therefore, several proposed solutions are discussed to address these limitations. An approach proposed by Zhou et al. (2022a) tries to break down reasoning into several sub-tasks. An alternative solution involves splitting the prompt into two distinct sections: the initial one, referred to as the “selection prompt,” would present relevant facts, while the subsequent section, known as the “inference prompt,” would involve drawing reasonings (Creswell et al., 2022).

2.6.4 Automatic Prompt

Automating prompts has emerged with the expansion of this discipline, aiming to generate prompts and improve the performance of LLMs on different tasks. Thus, several techniques were proposed to support this strategy. Automatic prompt approaches were developed by Shin et al. (2020) and Zhou et al. (2022b), automating the ways of generating prompts and then selecting appropriate ones.

A few studies introduce and discuss the concept of prompt pattern, which is important to achieve an efficient prompt engineering. Prompt patterns were developed to be a solid foundation for prompt engineering techniques (Schmidt et al., 2024). White et al. (2023) pointed out that prompt patterns were designed to be similar to software patterns, where software pattern is a method to provide a reusable solution to address a particular recurring issue in a particular context. Prompt patterns concentrate more on the interaction and produced results from LLMs, such as ChatGPT (White et al., 2023). The characteristic of prompt patterns is that they customise the result and interaction of LLMs.

2.6.5 Prompt Engineering for Extraction

Several studies have discussed the effectiveness of using prompt engineering techniques to enhance the quality of results generated by LLMs. Vijayan (2023); Polak and Morgan (2024)

confirm that designing questions and instructions (prompts) can improve the accuracy of data extraction. Prompt engineering techniques have been employed in various extraction tasks.

Polak and Morgan (2024) developed ChatExtract, a workflow that consists of a set of designed prompts to extract material properties, including Material, Value and Unit. This workflow is essentially based on the zero-shot prompting technique

Polat et al. (2024) conducted a study to adapt several prompt engineering techniques for effective knowledge extraction from text. The study focuses on chain-of-thought, self-consistency, and reasoning-and-acting techniques. These methods were evaluated using a relation extraction dataset (RED-FM). Although the approach works with this dataset and GPT-4, it cannot yet be extended to other models, as it was confined to specific datasets and models.

Tang et al. (2024) employed prompt engineering techniques to extract and summarise various medical information from the abstracts of research papers. This study focused on three types of prompt engineering techniques: chain-of-thought, few-shot prompting and persona. The evaluation was conducted using only two models: GPT-3.5 and GPT-4. However, the ground truth for this evaluation relied solely on human evaluators

Kanwal (2024) proposed BioREPS, a strategy to investigate and extract biomedical relations using chain-of-thought techniques and semantic similarity within a question-answering framework. The study evaluated the effect of different prompts on extracting biomedical relations using LLMs. However, decoder-only LLMs did not perform well for biomedical relation ion tasks.

Despite persistent efforts to combine prompt engineering with extraction tasks using LLMs, further improvement is needed for domain-specific extraction tasks. To the best of our knowledge, provenance extraction from publications remains an unexplored area, particularly in biomedical extraction tasks. Therefore, we conducted a prompt engineering experiment to address this gap, as explained in Section 3.4.2.

2.7 Summary

In this chapter, we introduced the main terminologies associated with dataset search, including the concept of datasets, dataset repositories and portals. Additionally, we explored the fundamental foundations of the dataset search domain, including various principles, definitions, process steps and existing taxonomies to enhance our understanding of its main functions. Furthermore, this study discussed the use of dataset search in biomedicine, including the techniques that aided biomedical researchers in searching for datasets, such as metadata. Moreover, we explained the

main foundations of provenance and its relationship to dataset search and biomedicine, including provenance models. We introduced LLMs, their use in biomedicine and techniques that aimed to improve LLMs, such as types of prompt engineering and its application to extraction tasks.

In this chapter, we highlighted the issues and difficulties encountered in biomedical domains when searching for datasets. These challenges included the difficulty of finding the history of datasets, which complicates reproducibility and trustworthiness, as well as concerns regarding data quality. In addition, [Gierend et al. \(2024\)](#) confirmed that there is significant heterogeneity in methods and models, as there are variations in dataset features, which presents several research opportunities in this field.

Therefore, in our research, we will first seek to confirm and understand the issues and difficulties encountered in biomedical domains during dataset search. We will assess how provenance can enhance dataset search from a biomedical perspective. Additionally, we will address this gap by integrating provenance into the dataset search process for biomedical researchers. Finally, we will evaluate the extent to which provenance contributes to the effectiveness of biomedical dataset search.

Chapter 3

Methodology

In this chapter, we present our research methodology to address the outlined research questions in Figure 3.1. Section 3.2 defines the dataset search requirements for biomedical researchers (RQ1). Section 3.3 describes the method for assessing the effectiveness of using provenance information in dataset search and identifying the specific provenance information elements needed (RQ2). Section 3.4 outlines the methods used to construct the provenance extraction process (RQ3). Finally, Section 3.5 explains the method for evaluating the enhancement of biomedical dataset search using provenance information (RQ4).

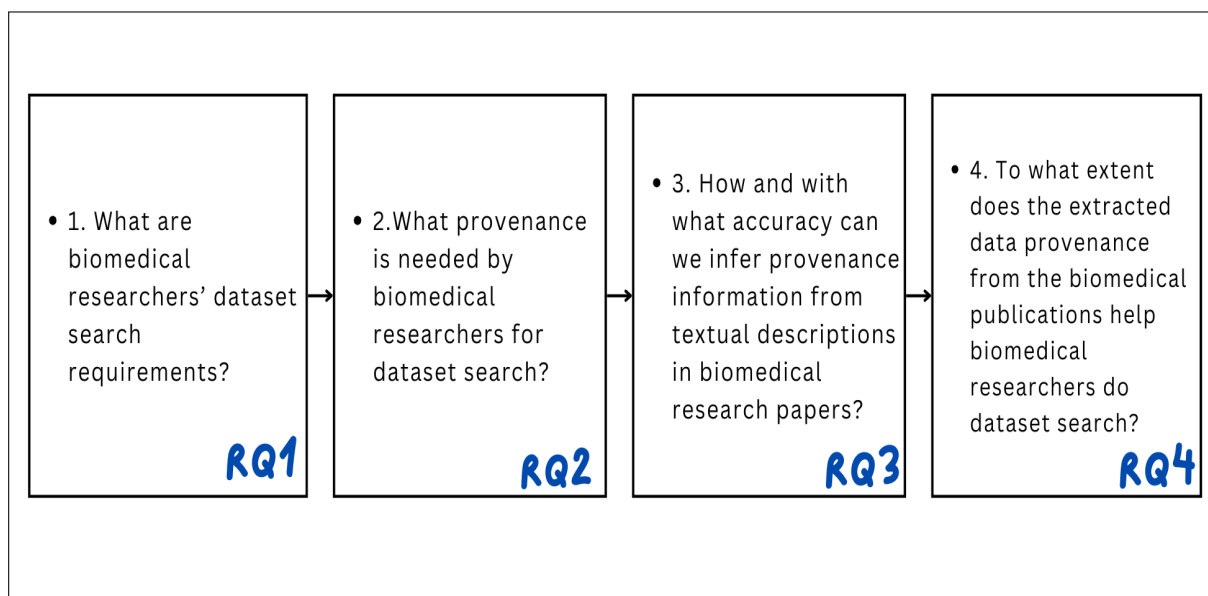


FIGURE 3.1: Research Questions (RQs)

3.1 Participant Recruitment Processes Overview

To answer the research questions outlined above, we conducted three human studies, which can be considered interrelated in terms of their recruitment processes. Several participants took part in more than one study. Figure 3.2 presents an overview of the participant cohorts and their overlap across the studies. In the first study (Section 3.2), we interviewed 13 participants (P1–P13), three of whom (P1, P3, and P9) also participated in subsequent studies. The second study (Section 3.3) was a large-scale evaluation involving 56 participants (P15–P63). This cohort included three participants (P1, P3, and P9) who had previously taken part in Study 1 and also participated in Study 3. The third human study (Section 3.5) involved 10 participants (P1, P3, P9, P14, P16, P17, P18, P19, P20, and P21). Of these, three participants (P1, P3, and P9) overlapped with both Studies 1 and 2, while the remaining seven had previously participated in Study 2.

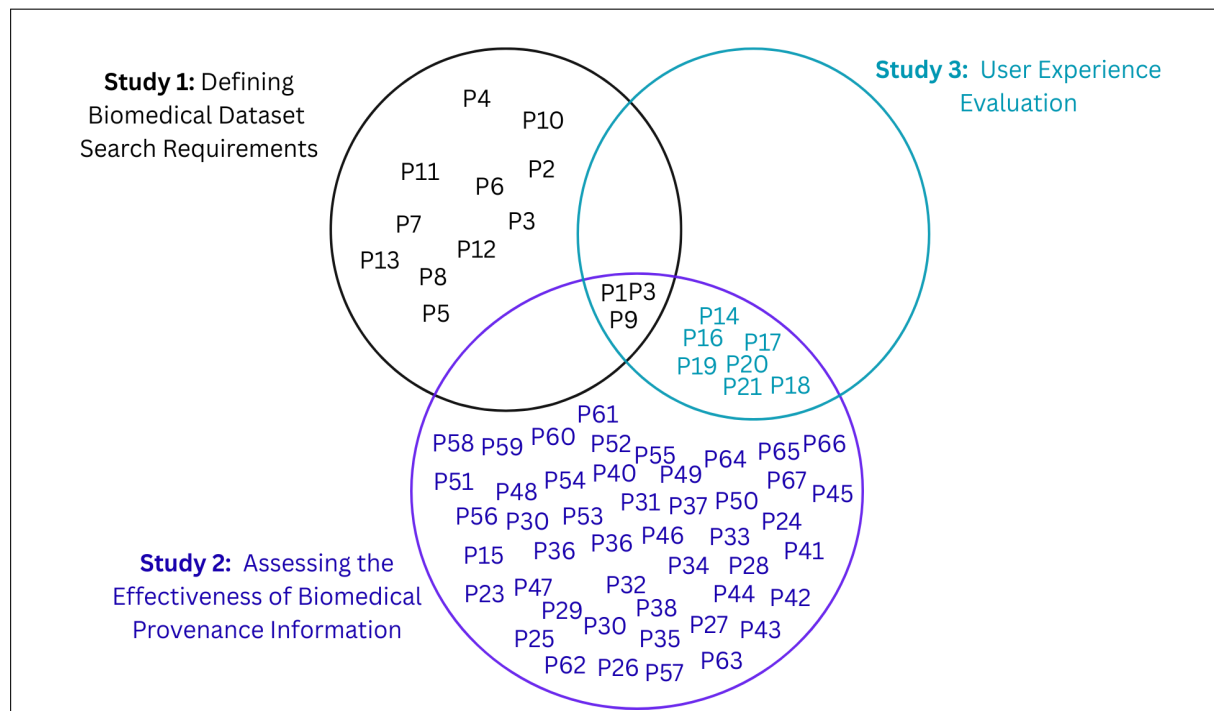


FIGURE 3.2: Participant cohorts

3.2 Methodology for Defining Biomedical Dataset Search Requirements (RQ1)

To address this question, a qualitative research method was chosen because it provides richer and more realistic data compared to other methods, enabling researchers to better understand

people's behaviours (Dudovskiy, 2016). One of the most effective ways to engage with biomedical researchers is through interviews, which allow us to explore their search methods and identify their specific requirements.

To collect qualitative data, semi-structured interviews are among the most suitable types to use (Dudovskiy, 2016). This approach can be used due to its ability to provide more authentic and comprehensive data compared to other methods. Additionally, it allows researchers to better understand people's behaviours. This type of interview allows for asking open-ended questions, followed by following-up questions to explore issues that may not have been identified previously (Cairns and Cox, 2008). Using a survey to address such this question may not provide a deep understanding, as it typically provides only superficial data (Dörnyei and Taguchi, 2009; Creswell and Creswell, 2017).

3.2.1 Semi-structured Interview design and development

Several objectives were identified for these interviews to answer RQ1. First, we aimed to gain an in-depth understanding of how biomedical researchers currently search the datasets and the challenges they face with these methods. Additionally, we sought to explore their requirements for improving dataset search techniques. The third objective was to support the common issues highlighted in previous studies. A complete list of the questions and their justifications can be found in Appendix A.

3.2.2 Interview procedures

The participants interviewed in this study were researchers working in the Faculty of Medicine (FoM) at the University of Southampton, who met our primary criteria: researchers involved in searching for or collecting datasets for their biomedical research projects. The initial participants were nominated by a colleague and invitations were sent via email. We then applied the exclusion criteria (see Figure 3.3) to finalise the participant list.

Following this, we employed a snowball sampling technique, where participants are asked to nominate other potential participants (Saunders et al., 2009). All potential participants were contacted via targeted emails, introducing the study and arranging interviews. Using this approach, a total of 43 researchers were invited to participate, resulting in 17 positive responses. Table 4.1 provides demographic information of the participants.

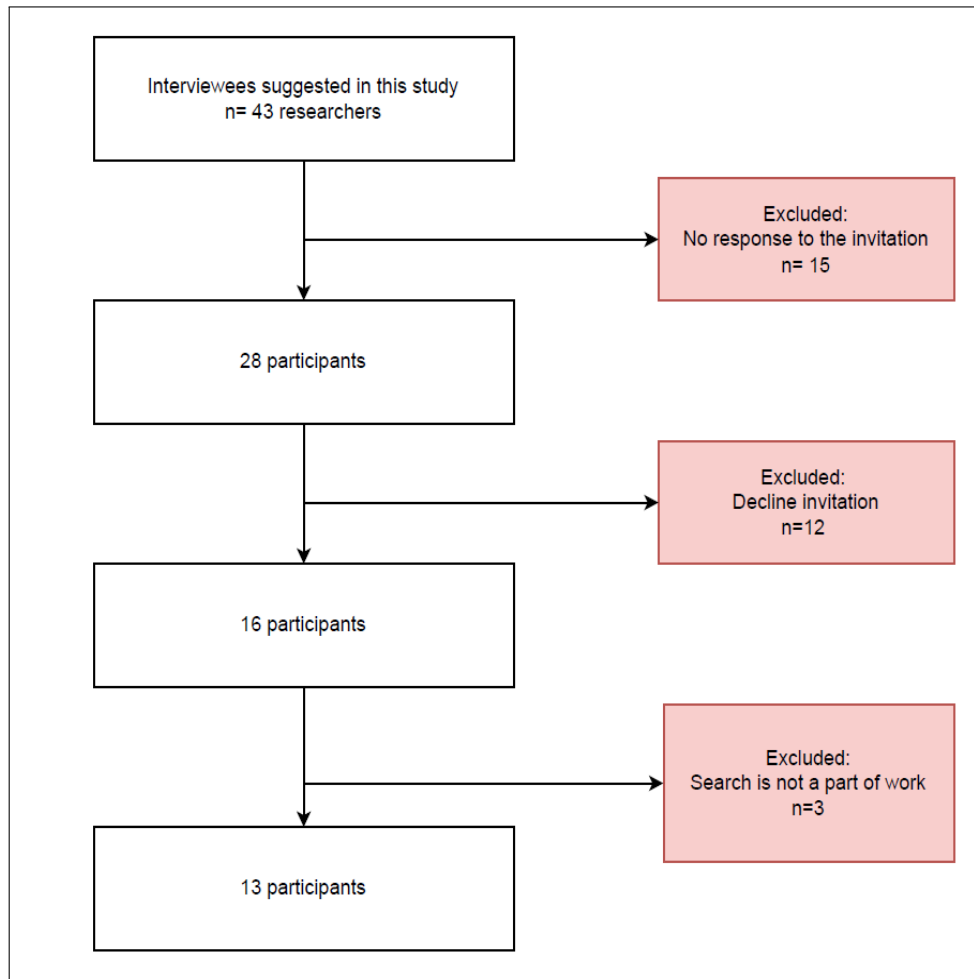


FIGURE 3.3: Exclusion criteria flowchart for Study 1

To determine an appropriate sample size, [Lilly \(1998, p. 64\)](#) states that a suitable number of interviews ranges between 5 and 25. Similarly, [Brinkmann and Kvale \(2018, p.43\)](#) confirm that most interview studies usually involve between 5 and 25 interviews due to several factors, including time constraints and available resources. The sample size is tied to the concept of saturation, meaning interviews are conducted until no new themes emerge and no further insights are gained ([O'reilly and Parker, 2013](#)). [Back \(2016, p.160\)](#) states that the “prevailing norm is to sample to theoretical saturation.” Similarly, [Hickman and Longman \(1994, p. 52\)](#) confirms that interviews are informative, with general insights typically obtainable through five to eight interviews, while seven to ten interviews may be sufficient to reach higher-level representatives and identify strategic conclusions. Although the initial goal of this study was to interview 25 participants, data collection was concluded once saturation had been achieved.

All interviews were carried out across a four-month period, from June to September 2022. Each interview lasted approximately 45 to 55 minutes. All participants agreed to have their interviews

recorded, except for one. Once the interviews were recorded, the audio files were transcribed for the analysis phase.

3.2.3 Interview Analysis

The analysis strategy used in this study was thematic analysis (Robson and McCartan, 2016), which is one of the most widely applied strategies in qualitative research. To conduct thematic analysis, Braun and Clarke (2006) recommend following six steps:

1. Familiarizing yourself with the data.
2. Generating initial codes.
3. Identifying themes.
4. Reviewing and refining these themes.
5. Defining and naming the themes.
6. Producing the final report.

In research methodologies, we used thematic analysis because of its flexibility, which can assist researchers in obtaining a rich, detailed understanding of the results (Braun and Clarke, 2012). Thematic analysis can be conducted using various approaches, including inductive and deductive methods (Braun and Clarke, 2021). The goal of inductive analysis is to derive themes or concepts through a detailed inspection of the raw data previously collected. In deductive analysis, the researcher seeks to confirm an existing hypothesis or assumption through the examination of empirical data. In our study, we employed an inductive approach, as our aim was to explore how biomedical researchers currently search for datasets and to identify their requirements to improve the dataset search process. In the coding process, we identified various main themes and sub-themes to extract deeper insights. We conducted the coding using NVivo12¹, a qualitative data analysis tool.

3.3 Methodology for Assessing the Effectiveness of Biomedical Provenance Information (RQ2)

We conducted a mixed-methods study, with data collected between June 2024 and August 2024. Employing mixed methods approach enables researchers to strengthen the analysis and

¹<https://lumivivo.com/products/nvivo/>

evaluation of the collected data (Sandelowski, 2000; Creswell and Clark, 2017). In this study, we used quantitative and qualitative data collection techniques to examine the effectiveness of using provenance information in dataset search and to identify the specific provenance information elements needed.

3.3.1 Identifying critical provenance information for biomedical researchers

Building on existing works in the literature (Johns et al., 2023), which identified the provenance needs in the biomedical research domain as well as semi-structured interviews conducted with biomedical researchers to determine dataset search requirements (4.2.8), we developed a visual representation diagram to present the necessary provenance information for enhancing dataset search. Johns et al. (2023) highlighted the high level of the provenance aspects needed in biomedical research domain, which is very compatible with our findings mentioned in Table 4.2. Table 4.3 summarises the provenance requirements. This provenance information encompasses all details of the experiments, from obtaining ethical approval to storing the outcomes on which the dataset is based.

Since the provenance information is initially unstructured, we manually encoded it using the PROV-DM W3C recommendation into activities, such as patient recruitment; entities, including the tools used; agents, including all researchers involved in the experiment; relations, including all relationships between these previous concepts. This information was then presented using a visual representation method, adopting an activity-centred design in which other information was linked through relationships, as illustrated in the example shown in Figure 3.6.

An activity-centered graph is a method used to present a sequence of activities, displaying a workflow from the first step to the last, while detailing all components involved in the progression of activities (Goodyear et al., 2021). Activity-centered design is considered one of the best approaches for understanding both physically and socially situated contexts, as it can present several components, such as material artifacts and digital tools (Goodyear and Carvalho, 2016). Therefore, we selected this approach to present provenance using PROV-DM. This involves placing activities at the centre, with entities and agents positioned around them, all linked by relationships.

3.3.2 Survey Design and development

An online questionnaire was utilised to answer our research questions. According to Cohen et al. (2007), collecting needed data through a questionnaire can be more reliable since respondents

can complete it privately. We developed an online questionnaire using Qualtrics². The design of a questionnaire depends on the research questions or goals being addressed. Without conducting a previous design, there may be an opportunity to collect irrelevant information, potentially altering the direction of the study (Boynton and Greenhalgh, 2004). Additionally, the accuracy and quality of the responses can be affected by the survey design (Brace, 2018). Therefore, we followed a structured approach to develop a well-designed questionnaire (Jenn, 2006), which involved the following steps:

1. Constructing a conceptual framework: researchers should have a clear understanding of the research questions.
2. Designing the right questions: researchers need to identify the appropriate question types, such as open-ended questions.
3. Ensuring comprehensive answer choices: the options provided for each question should be exhaustive.
4. Using filtering techniques: This guides participants to skip questions that may not apply to them.
5. Ordering questions logically: All questions should follow a logical sequence.
6. Pre-testing or pilot testing: Conducting a test before distributing the final questionnaire to ensure the reliability and clarity.

The questionnaire is structured around the research questions and is divided into four main sections. The first section includes the consent form and the participant information sheet link. The participant must provide consent in order to proceed to the second section of the survey. The second section collects demographic details, including the participants' research areas, and the seniority of their roles, as presented in Section 5.1. The third and fourth sections present two dataset search tasks, which are detailed explained in the following section, aimed at collecting data to assess the effectiveness of using provenance information in selecting datasets and evaluate the presented provenance information elements. Two dataset search scenarios were presented, with each scenario involving different types of metadata, such as provenance information metadata.

3.3.3 Dataset search tasks

After collecting demographic details, we designed two imagination tasks involving the biomedical dataset search online. Imagination is the process of simulating an action or forming a mental

²<https://www.qualtrics.com/>

image (Jung et al., 2016; Markman et al., 2012). Furthermore, Beese et al. (2019) confirms that the use of simulation tasks is important in the field of information systems. As a case study, Zhang et al. (2023b) used an imagination task in a series of exploratory studies for dataset expansion.

To design these tasks, we sought assistance from an independent expert in the biomedical domain to help design the tasks following an iterative process. In the first iteration, we established that a common information need is the identification of factors related to a disease, including causes, risks and contributing factors, with the latter being often an information need on its own. Based on their advice, we designed two search tasks: one focusing on contributing factors and the other on causes and risks. In the second iteration, the expert was asked to assess the understandability of a task until satisfaction they were close enough to a real dataset search tasks and are understandable to biomedical researchers at different levels. In the third iteration, we asked the expert to conduct a search in public repositories and a literature search engine to identify datasets relevant to the tasks to be used in the study, as these are common places for biomedical dataset search, as observed in the findings of Chapter 4.

To ensure each task involved different relevant datasets, we decided to focus on different diseases:

The first task asked our participants to imagine a scenario where they need to search for one or more datasets to conduct research on Inflammatory Bowel Disease:

“Task 1: Imagine you are a new postgraduate biomedical researcher. You would like to initiate a project looking at factors contributing to Inflammatory Bowel Disease (IBD) in infants. You need a dataset for investigation purposes to understand this disease. As you search, you need to identify whether a particular dataset is worth downloading, extracting and evaluating for your use.”

The second task included a scenario to search for datasets to investigate the factors influencing Crohn’s disease:

“Task 2: Imagine you are a research fellow in a biomedical research group. One part of this group is focused on Crohn’s disease, including its history, causes and possible risks. You would like to initiate a project looking at factors contributing to Crohn’s disease. Your task requires a dataset to investigate the impact factors of this disease. As you search for a dataset, you need to identify whether a particular dataset is worth downloading, extracting and analysing for fitness for your use.”

To provide further context of the cost of downloading a dataset, we added the following notice after the description of each task “Note: Downloading this dataset for in-depth inspection will take 90 minutes of your time”.

Following that, the same types of the dataset metadata options were provided in a different order to explore whether any difference when changing the order. Under these dataset search tasks, four dataset options were provided for participants to select from:

Option A) The first option was basic metadata of the dataset typically available in dataset repositories. We showed the metadata from Datamed ³ of the relevant dataset. As mentioned earlier, it uses the DATS model, designed to collect information about dataset elements, including experimental datasets and other related elements, such as publications and data types (Sansone et al., 2017).

A keyword search was conducted using “Inflammatory Bowel Disease,” that returned 2,055 results. The volunteer expert selected one dataset relevant to the task. The selected dataset includes several elements: the dataset name, repository, identifier, data or study type, source organization, access conditions and an access hyperlink. Figure 3.4 presents a sample of metadata taken from Datamed. The first question following the metadata was presented whether the participant would download the dataset based on the provided metadata.

Typically, this metadata includes several brief elements to describe existing datasets, entered by the dataset holder, and is sometimes incomplete. A user needs analysis study conducted by Dixit et al. (2018) emphasised that some dataset searchers faced difficulties in evaluating the relevance of datasets retrieved by DataMed due to incomplete and low-quality metadata.

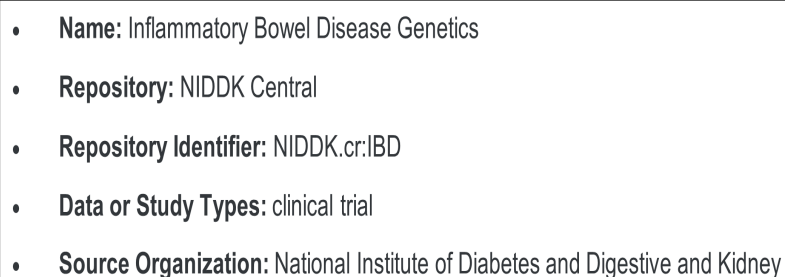
- 
- **Name:** Inflammatory Bowel Disease Genetics
 - **Repository:** NIDDK Central
 - **Repository Identifier:** NIDDK.cr:IBD
 - **Data or Study Types:** clinical trial
 - **Source Organization:** National Institute of Diabetes and Digestive and Kidney

FIGURE 3.4: An example of metadata from Datamed.

Option B) The second option was an abstract of the article published based on the dataset sourced from PubMed ⁴. PubMed provides a list of results based on an entered query, with each result containing only the title and/or abstract (Canese and Weis, 2013). In bioinformatics, datasets are

³<https://datamed.org/>

⁴pubmed.ncbi.nlm.nih.gov/

often described in academic publications, as identified in the findings of the first study detailed in Chapter 4. Therefore, we used the abstract of the publication related to the dataset we chose to the task.

We conducted the same search technique using the same keywords, which resulted in over 136,000 options. Figure 3.5 presents a sample of metadata taken from Pubmed, which includes the title and abstract.

Title: Elevated Levels of the Cytokine LIGHT in Pediatric Crohn's Disease
Abstract: LIGHT (homologous to lymphotoxins, exhibits inducible expression, and competes with HSV glycoprotein D for herpes virus entry mediator, a receptor expressed by T lymphocytes), encoded by the TNFSF14 gene, is a cytokine belonging to the TNF superfamily. On binding to its receptors, herpes virus entry mediator and lymphotoxin β receptor, it activates inflammatory responses. We conducted this study to determine whether plasma LIGHT levels are elevated in Crohn's disease (CD) in a pediatric population with the aim of nominating this cytokine as a therapeutic target. We used a single-molecule immunoassay to determine the circulating levels of free LIGHT in plasma from pediatric patients with CD in our biobank (n = 183), a panel of healthy pediatric (n = 9) or adult (n = 22) reference samples, and pediatric biobank controls (n = 19). We performed correlational analyses between LIGHT levels and the clinical characteristics of the CD cohort, including age, Montreal classification, family history, medical/surgical therapy, and routine blood test parameters. LIGHT levels were greatly elevated in CD, with an average of 305 versus 32.4 pg/ml for controls from the biobank ($p < 0.0001$). The outside reference samples showed levels of 57 pg/ml in pediatric controls and 55 pg/ml in adults ($p < 0.0001$). We found a statistically significant correlation between white blood cell count and free LIGHT ($p < 0.046$). We conclude that free,

FIGURE 3.5: Abstract on Pubmed for paper (Cardinale et al., 2023)

Option C) The third type of dataset metadata provided was provenance information, which is central to our contribution. This provenance information was extracted from a published paper related to the dataset on PubMed. One published article was selected by the same volunteer expert from the previous option. All provenance information was extracted, including the procedures, steps and tools used in the experiments.

The aim of this task is to evaluate the understanding of provenance metadata presented in a manually created diagram. This diagram serves as a visual representation to convey research procedures and data processing steps, as explained in Section 2.3.3. Since the extracted provenance information was unstructured, it was manually organised using PROV-DM into activities, entities, agents, and relationships, as described earlier in Section 3.3.1, and subsequently presented in the visual representation. An example of this design is shown in Figure 3.6.

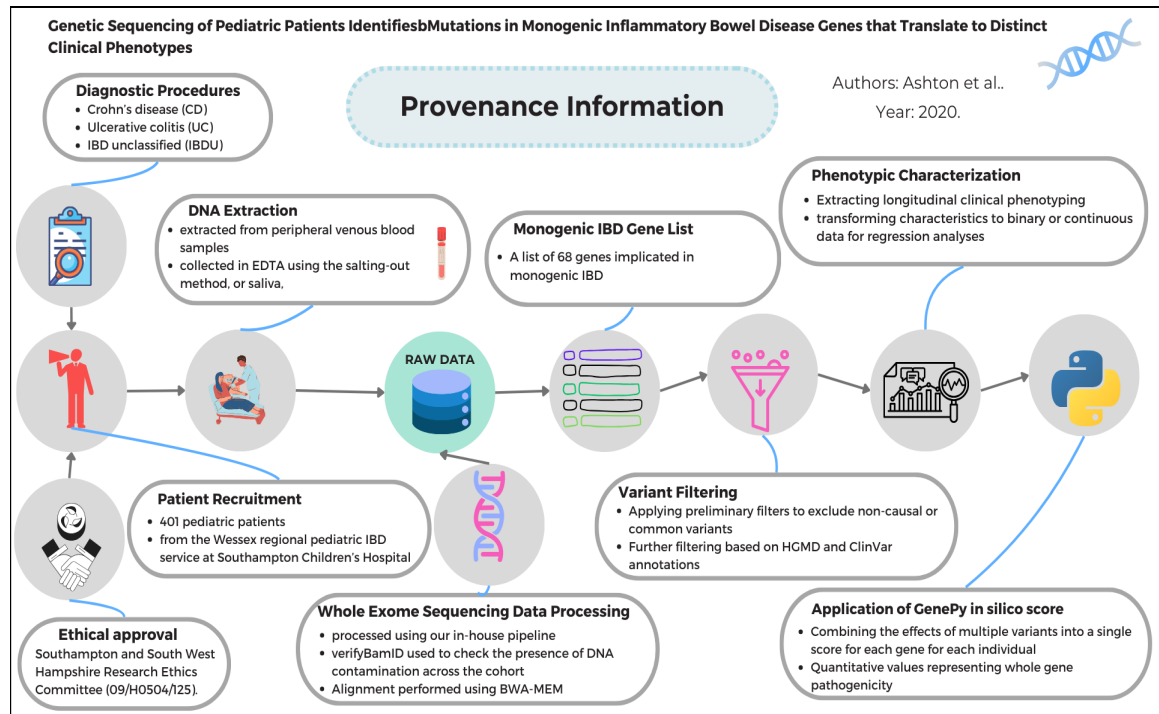


FIGURE 3.6: Provenance information arranged as an activity-centered design.

Option D) The final type of dataset metadata option provided was presenting **both** (B) abstract of a published paper and (C) provenance information with a visual representation. The purpose of this options to discover if this combination is better than options (B) or (C) alone. This type of dataset description includes a detailed summary of the published information along with the extracted provenance information from the article. In short, it is an integration of the second and third types from this task.

Below each of option, we first asked a closed-ended question of whether they would download the dataset or not based on the presented information. If our participants chose to download it, then they were asked to assess the usefulness of the provided information using a five-point Likert scale (example in Figure 3.7). If they chose not to download the dataset, they were asked to elaborate on their reasons for not selecting the dataset in an open-ended text field. To explain further, if a participant decided to download the dataset and assess the usefulness of the provided information without reporting any issues, we assumed the metadata option was good for helping assess relevance. Otherwise, they were asked to provide reasons that led them to continue searching for more datasets.

To understand what useful information participants get from the provenance information, we asked those participants that chose “Option C: Provenance information with a visual representation” or “Option D: Abstract and Provenance” as their preferred presentation, to select one or more pieces of information about the dataset they considered they gained. We provided

multiple-response question, which includes six options. The options were presented as follows: (1) Was ethical approval information acquired?; (2) How the data has been collected; (3) How the data has been processed; (4) Understanding the steps followed; (5) Understanding the outcome of this experiment; (6) Other (input field for participant to add). Our choices were motivated by the findings from the interviews conducted in the first study (4.2.8), and they are also consistent with the key aspects of provenance reported by (Johns et al., 2023) and (Samuel and König-Ries, 2017). The whole questionnaire consisted of 28 questions, plus three demographic questions.

Q2: If yes, what part of the above information helped you decide?
 Tip: Rank each element from 0 (Not useful at all) to 4 (Very useful)

0 (Not useful at all) 2 (Somewhat useful) 4 (Very useful)

0 4

Title

Abstract

Q2: If yes, what part of the above information helped you decide?
 Tip: Rank each element from 0 (Not useful at all) to 4 (Very useful)

0 (Not useful at all) 2 (Somewhat useful) 4 (Very useful)

0 4

Title + Authors + Year

Provenance Information Details

Relationships between the steps

Explanatory Icons

FIGURE 3.7: Examples of the five-point Likert scale used for the usefulness assessment.

3.3.4 Survey Validation

After completing the questionnaire design, an independent expert in the field of Human-Computer Interaction was consulted to review the structure of the survey, focusing on the presentation of information and ease of use. Without examining the design structure, there is a risk of collecting irrelevant data, which could potentially alter the direction of the study (Boynton and Greenhalgh, 2004). Additionally, the accuracy and quality of the responses can be affected by the survey design (Brace, 2018).

Following the prior review of the questionnaire design, pilot tests were conducted to validate both the questions and the information presented in the scenarios. According to Sharp et al.

(2019), revising and piloting the survey is a crucial step before its application. In this process, seven individuals were involved, including four professionals and three postgraduate researchers at the University of Southampton, UK.

3.3.5 Survey procedures

The questionnaire was distributed through two main distribution methods. First, promotion at a local conference focused on Omics data, including genomics, proteomics, metabolomics and transcriptomics held in July 2024. The approximate number of attendees at the conference was over 100. Second, sending a recruitment email through two mailing lists specialised in biomedical fields at the University of Southampton.

In this study, it was not possible to precisely determine the exact number of biomedical researchers. Therefore, we explored common methods for estimating sample size in both quantitative and qualitative research. One widely used method is applying statistical tests for sample size estimation, which can be performed using G*Power (Faul et al., 2009). This software is commonly used for sample size estimation and power calculations. Additionally, Kang (2021) mentioned that G*Power enables researchers to determine sample size and conduct power analysis. This process involves several factors, including effect size, power ($1 - \beta$) and the type of statistical analysis to be conducted (Kang, 2021).

To estimate the minimum required sample size using the G*Power tool, we applied the following parameters: an effect size of 0.5, an expected power of 0.95 and a t-test for analysis, as illustrated in Figure 3.8. As noted by Cochran (1963); Mayr et al. (2007), estimating sample size with G*Power depends on the specified effect size. In this case, an effect size of 0.5 was utilised, resulting in a recommended sample size of 54 biomedical researchers.

Initially, 156 participants agreed to take part and began filling out the questionnaire. However, due to the importance of most questions, all incomplete responses were excluded from the analysis. Therefore, only 56 responses were considered in the data analysis phase, representing 35.89% of the total responses. The primary reason for the incomplete responses might be the length of the questionnaire, which took approximately 20 minutes to complete, whereas the ideal length is around 10 minutes (Revilla and Höhne, 2020).

The analysis for this study was conducted using both quantitative and qualitative techniques, as the collected data includes numerical and descriptive information. For the quantitative data, SPSS software was used to perform several statistical analyses and tests. Descriptive statistics were conducted to summarise key variables, including the mean, median and standard deviations.

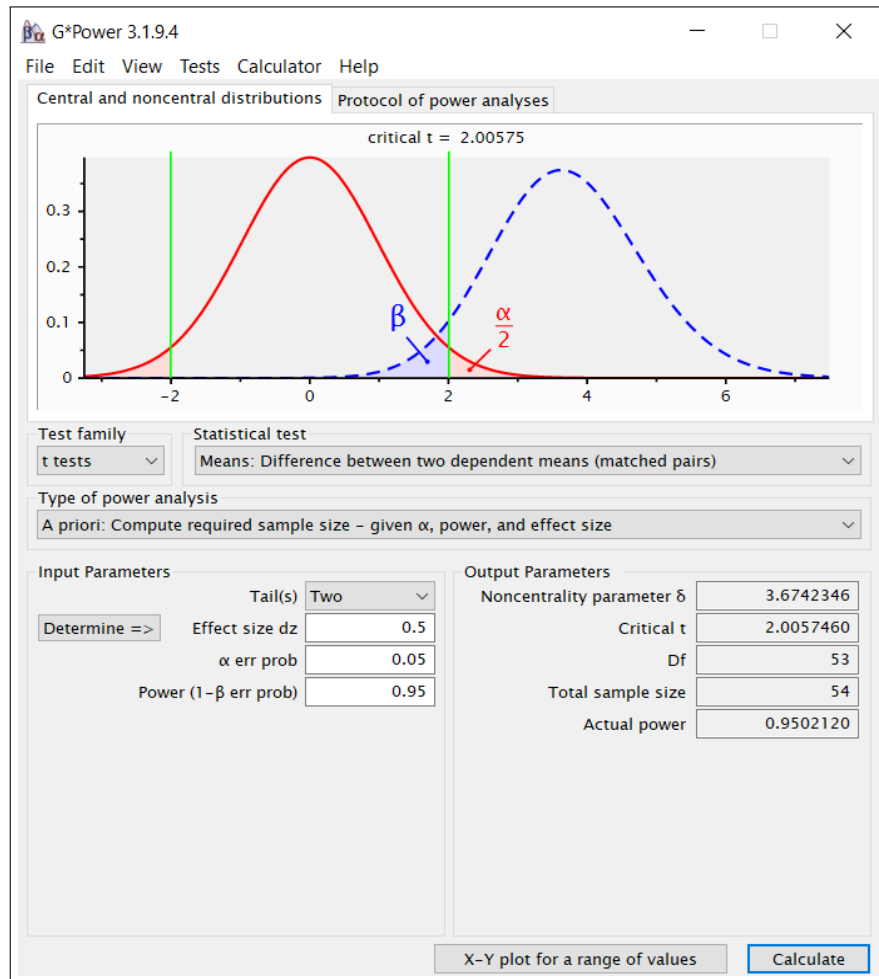


FIGURE 3.8: G*Power for sample size estimation.

These statistics aim to describe the features or attributes of a given sample by representing “the midpoint of a range of scores” (Fisher and Marshall, 2009, p.93).

For examining the relationship between participants’ roles and their selection of dataset metadata options, Chi-Square test were performed. This test aims to examine the relationship between two categorical variables or to evaluate how well a distribution fits a group or population (Franke et al., 2012).

Several independent t-tests were also conducted to compare the means of the existing dataset metadata approaches and the proposed ones. The independent t-test aims to compare the means of two independent groups (Okoye and Hosseini, 2024). To assess differences among multiple dataset metadata approaches, an ANOVA analysis was applied. ANOVA is used to compare the means of more than two groups (Kim, 2014).

Finally, to ensure the consistency and accuracy of the five-point Likert scales used in the questionnaire, various reliability tests were performed. Reliability tests aim to measure the

consistency of scale values (DeCoster and Claypool, 2004).

For the qualitative data collected, we used NVivo12 to analyse all open-ended questions. This tool is designed to support various data types and formats, enhancing the overall quality of analysis (Dhakal, 2022). The results of these tests and the qualitative thematic analysis will be detailed in Chapter 5.2.

3.4 Extracting provenance automatically (RQ3)

To address this research question, we first developed an automated provenance extractor using an LLM, as explained earlier in detail in Section 2.5. Additionally, we designed several prompts to extract provenance information from publications, aiming to enhance the extraction process using ChatGPT-4o (version 2025-02).

3.4.1 Extraction Tasks using LLMs

There is significant interest in using LLMs for information extraction, as they have shown promising capabilities in several NLP tasks. Several studies Xu et al. (2023); Liu et al. (2024b); Peng et al. (2024); Zhang et al. (2024) confirm notable advancements in NLP domains with the advent of LLMs, which exhibit exceptional abilities in understanding and generating text. Additionally, LLMs can perform complex tasks, leading to successful task completion (Bender et al., 2021). Furthermore, LLMs are making significant progress in IE tasks (Bommasani et al., 2021; Goel et al., 2023). Moreover, LLMs have demonstrated considerable capabilities in various biomedical domain tasks (Li et al., 2023; Thirunavukarasu et al., 2023; Zhou et al., 2024).

We constructed our provenance extractor from biomedical publications using LLMs. This was achieved using GPT-4o via the OpenAI API. This model was selected for several reasons. As of February 2025, GPT-4o became available to web users with limited access; however, full access requires a ChatGPT Plus subscription. GPT-4o is distinguished by its ability to process and generate text, images, audio, and video. It is also significantly faster than previous versions, including GPT-4-turbo and GPT-3.5. ChatGPT-4o excels in its capability to understand and produce information across several domains, making it well-suited for applications like customer assistance and content creation (Ekin, 2023). The respondents' quality of ChatGPT was assessed in several domains, for instance, answering medical licencing exams (Gilson et al., 2022) and mathematical reasoning (Frieder et al., 2023).

In addition, we employed various techniques for processing, formatting, extracting, and storing data to build the extractor. The architecture of the extractor is discussed in detail in Section 6.1.

To enhance the performance of LLMs for specific tasks and achieve optimal results, prompt engineering has emerged as a significant research topic and an essential skill for users (Meskó, 2023; White et al., 2023), as discussed in Section 2.6. Numerous studies highlight the effectiveness of using prompts in retrieving information from LLMs (Reynolds and McDonell, 2021; Schick and Schütze, 2021).

3.4.2 Provenance Extraction Prompts for Biomedical Research

A few studies introduce and discuss the concept of prompt pattern, which is important to achieve an efficient prompt engineering. Prompt patterns were developed to be a solid foundation for prompt engineering techniques (Schmidt et al., 2024). White et al. (2023) pointed out that prompt patterns were designed to be similar to software patterns, where software pattern is a method to provide a reusable solution to address a particular recurring issue in a particular context. Prompt patterns concentrate more on the interaction and produced results from LLMs, such as ChatGPT (White et al., 2023). The characteristic of prompt patterns is that they customise the result and interaction of LLMs.

Despite several current efforts to understand the efficiency of prompt concepts in LLMs (Wang et al., 2022), there remains a gap in designing and refining specific prompts for IE in general, and particularly for the extraction of provenance information from scientific papers. Therefore, we designed eight prompts specifically for this task and subsequently evaluated them in Section 6.2 to identify the most effective one.

This section focuses on the proposed prompts for extracting provenance information from publications. The aim is to achieve better output from LLMs, potentially at a lower cost (at the time). All the following prompt patterns were presented and discussed in (White et al., 2023). We adapted these patterns to suit our purpose and subsequently evaluated them. The prompt patterns that are recommended are provided together with their objectives and contextual statements:

- **Persona Pattern (A):** The Persona Pattern is a prominent pattern within the Output Customisation category (White et al., 2023). The aim of this pattern is to guide and refine the output of the LLM. By providing an LLM with a specific a “persona”, it may become more skilled in determining its replies to be more relevant and specific. For

example, we may ask the LLM to act as a software engineer when we need it to review a code error. For instance, when seeking code error reviews, we could ask the LLM to act as a software engineer.

The proposed statement for the provenance extraction is: “Act as a provenance expert, (...) Can you please extract all provenance information of ?”

- **Recipe Pattern (B):** This prompt pattern enables users to obtain a series or sequence of steps or actions to achieve a specific outcome, sometimes with certain knowledge, constraints or instructions. It could be beneficial when utilising this prompt to firstly provide LLM with the primary aim of the task. For example, if we require a summary of the findings of a particular study, the following statement may be used to delineate our purpose: “I am endeavouring to summarise the outcome of (...)”. Next, we suggest outlining the list of steps required to be encompassed in the final result. Subsequently, we would direct the LLM to provide a complete sequence of steps.

The contextual statement suggested for provenance information extraction is: “I am trying to extract the provenance information from (.....) paper, which is titled (.....), I know that I need to identify the activities, entities, agents and relationships between them for the implementation. Please provide a complete sequence of all activities, entities, agents and the relationships between them that are used (.....), which would help (.....) in the experiment’s reproducibility.”

- **Question Refinement Pattern (C):** The purpose of this pattern is to ensure that the conversational language model consistently proposes alternative questions that are potentially superior or more refined than the user’s initial question. Providing the scope or domain area of the prompt, for instance biomedicine, would assist the LLM in not deviating from the primary field whilst refining the question.

The proposed statement for the provenance extraction is: “Within the scope of provenance, suggest a better version of the question that can be used to extract all provenance information, including activities, entities, agents and the relationships between them of the experiment implementation from the paper titled (...)”.

- **Alternative Approaches Pattern (D):** Using the alternative approaches pattern enables LLMs to propose various approaches for achieving a task. By utilising this pattern, users may be presented with additional approaches that exceed their current level of comfort.

The contextual statement for provenance extraction is: “Within the scope of extracting provenance information from (...) paper, which is titled (...), if there are alternative ways to accomplish the same thing with the same paper. List the best alternative approaches.”

- **Cognitive Verifier Pattern (E):** Studies in academic literature have shown that LLMs tend to have improved reasoning abilities when a question is broken down into several sub-questions, each providing responses that collectively provide the overall solution to the original issue.

The purpose of the pattern is to encourage the LLM to consistently split the main question into sub-questions to be answered by the user before integrating those answers to generate the final response for the overall question. The proposed statement for provenance extraction using LLMs is: “When I ask you a question regarding provenance information extraction from a (.....) paper, generate three additional questions that would help you give a more accurate final answer. When I answer these questions, combine the answers to provide the final answer to my original question.”

- **Experimenting with context and example Pattern (F):** The use of this prompt pattern allows users to guide LLMs by providing contextual information or presenting examples in the prompt input. This pattern can enable the LLM to analyse given contexts or examples and predict more accurate and relevant responses. The recommended statement for the extraction of provenance is as follows: “I will provide you with a (...) paper. I need you to read the whole paper and extract all provenance information from the experiment. For instance, (...) can be classified as an action, (...) an entity, and (...) as an agent. The aim of this task is to use this provenance information to help biomedical researchers understand and reproduce this experiment.”
- **Scenario Pattern (G):** The use of this prompt pattern allows users to provide more deliberate or analytical details. It can help the LLMs provide a more thoughtful and detailed answer to the prompt. The recommended statement for the extraction of provenance is as follows: “Imagine a scenario where you would like to know how a wet-lab experiment was implemented, what was the data used in the experiment, all techniques in the lab, all used tools, etc., in order to reproduce the experiment. You have only an opportunity to access a published paper regarding this experiment. Thus, you need to read the whole paper and extract all the provenance information of the experiment, including all activities, entities, agents, and relationships between them, based on the PROV-DM. This paper is titled (...).”
- **Identifying the Overall Goal of the Prompt First Pattern (H):** This type of pattern aims to characterise the kind of output being sought from LLMs. It would help LLMs provide more relevant responses with a more structured style. The contextual statement suggested for provenance information extraction is: “I would like you to extract the provenance information of a wet-lab experiment from the given article, which is titled: (...). Adapt

your answer based on PROV-DM components, which means I want you to extract all activities of the experiments, all used entities, and all provided agents. Can you identify all relationships between the components as well?”

3.4.3 Provenance Extraction Experiment

To investigate the capability of ChatGPT-4o in extracting provenance information from publications and to assess the most effective prompt patterns, we conducted the following experiment.

3.4.3.1 Collecting a set of papers from biomedical researchers

We identified several biomedical researchers from a previous study (Section, 3.2.1) to review their publications and select papers to be used for this evaluation. Google Scholar and PubMed were used as the primary search engines to locate articles by their names. In this search process, our initial focus was on the title and abstract of each article. As a result, we initially downloaded 21 papers. These papers were then organized in Microsoft Excel, containing the paper title, year, authors, abstract, introduction and conclusion. We applied several inclusion criteria:

- Articles published by six local experts in “Biomedicine” who are not involved in this project but are available for domain-specific assistance.
- Articles based on wet-lab “Omics” experiments.
- Articles published within the last five years (2018-2023).
- Articles with varying lengths and formats.

After applying the inclusion criteria, we selected six papers to conduct the prompt evaluation experiment, which are presented in Table 3.1.

3.4.3.2 Systematic Approach

We designed a systematic approach to examine the prompts using an LLM and select the most effective prompts. This approach involves four main steps, as shown in Figure 3.9: **The first step** is intended to ensure that the LLM possesses full access to the entire contents of the paper after uploading it. During this step, we initially ask the LLM several questions about the number of main sections, references, and pages in the article to guarantee that the model can analyse,

ID	Title	Citation
1	Faecal virome transplantation decreases symptoms of type 2 diabetes and obesity in a murine model	(Rasmussen et al., 2020)
2	Genetic Sequencing of Pediatric Patients Identifies Mutations in Monogenic Inflammatory Bowel Disease Genes that Translate to Distinct Clinical Phenotypes	(Ashton et al., 2020)
3	Prediction of Crohn's Disease Stricturing Phenotype Using a NOD2-derived Genomic Biomarker	(Ashton et al., 2023)
4	Immunological Profiling of Paediatric Inflammatory Bowel Disease Using Unsupervised Machine Learning	(Coelho et al., 2020)
5	Blood gene expression predicts intensive care unit admission in hospitalized patients with COVID-19	(Penrice-Randal et al., 2022)
6	Analysis of Mutation and Loss of Heterozygosity by Whole-Exome Sequencing Yields Insights into Pseudomyxoma Peritonei	(Pengelly et al., 2018)

TABLE 3.1: List of Biomedical Articles

scan and read the whole content. **The second step** is to emphasise that the LLM understands the concept of provenance as well as the PROV-DM (Missier et al., 2013a). Implementing this step would improve ChatGPT's comprehension of the task's context and enable it to tailor its response to the PROV-DM. Once the LLM has thoroughly accessed the whole content and showed a comprehensive understanding of PROV-DM, **the third step** aims to pose the provenance extraction prompt to ChatGPT-4. For each iteration of this experiment a different prompt from Section 3.4.2 was applied to each of the 6 papers (Table: 3.1). The prompt should be complete and grammatically correct prompt. **The final step** is to ask the model to identify the relationships between all the retrieved components from the previous step based on the PROV-DM. Each prompt pattern was examined in an individual session due to the fact that each session possesses its own short-term memory, as confirmed by the ChatGPT's support team.

3.4.3.3 Metrics for Evaluating Prompts

To evaluate provenance extraction using a LLM, we first designed the systematic approach involving a series of interaction steps, which is explained in detail in Section 3.4.3.2. We then applied this approach to test ChatGPT-4 using six research papers. Following this, we developed a scoring scale for intermediate prompts. The scoring scale was defined as follows:

- 2 if the prompt answered correctly on the first time.
- 1 if the prompt answered correctly in multiple attempts (up to 5 attempts).
- 0 if no prompt answered correctly in multiple attempts (up to 5 attempts).

We sought assistance from an independent biomedical expert to extract provenance information in PROV-DM for each of the six papers described above, which served as the ground truth for this

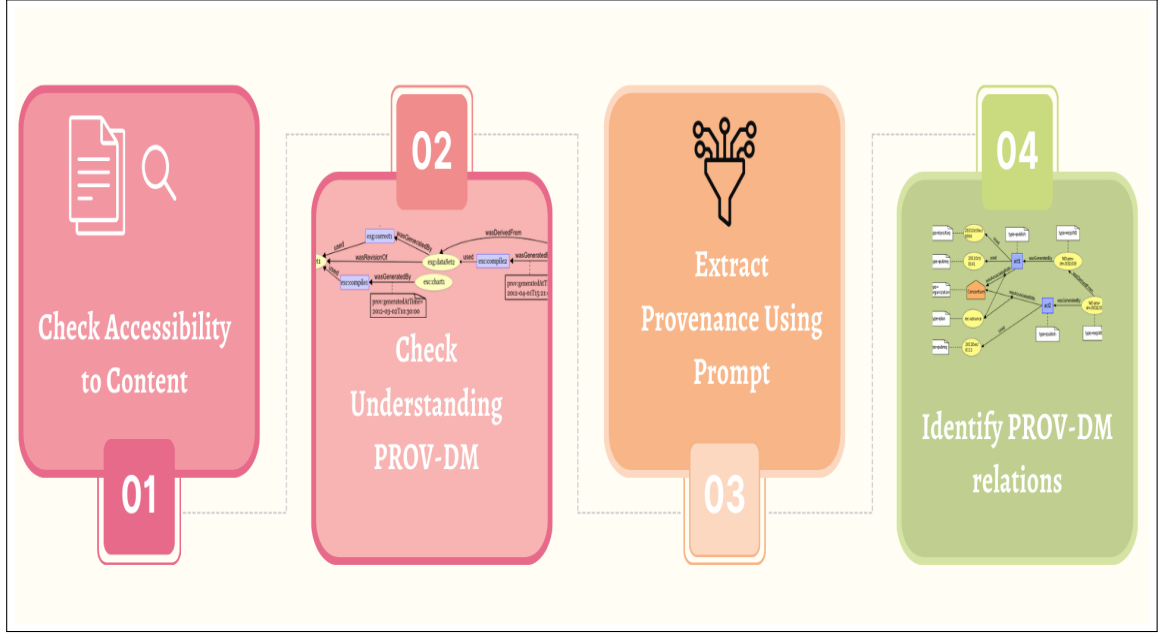


FIGURE 3.9: Four interaction steps followed with the LLM.

evaluation. We then compared this ground truth with the provenance information generated by the LLM. We also established traditional information retrieval metrics to evaluate the effectiveness of the provenance extraction prompts that are **Precision** and **Recall**. The results of this experiment will be presented and discussed in Section 6.2.

The formula of **Precision** is

$$Precision = \frac{|P \cap Pe|}{|P|} = \frac{\text{Fraction of relevant provenance information from the LLM}}{\text{All retrieved provenance information from the LLM}} \quad (3.1)$$

where the provenance information can be entities, activities and agents. There is provenance information returned from the LLM (p) which contains $\{e, a, c, r\}$ and provenance from an expert (Pe) which contains $\{ee, aa, ce, re\}$.

The formula of **Recall** is

$$Recall = \frac{|P \cap Pe|}{|Pe|} = \frac{\text{Fraction of relevant provenance information from the LLM}}{\text{All relevant provenance information from the ground-truth}} \quad (3.2)$$

where the provenance information can be entities, activities and agents, in which there is provenance information returned from the LLM (p) which contains $\{e, a, c, r\}$ and provenance information from an expert (Pe) which contains $\{ee, aa, ce, re\}$.

3.4.4 Scalability Experiment

To evaluate the performance and scalability of any system or application, it is crucial to conduct performance and scalability testing or software estimation (Al-Said Ahmad and Andras, 2019; Nasir, 2006). The purpose of this test is to evaluate several dimensions: performance under different load conditions, performance with scaling various resources and performance under varying volumes and stress levels (Iyer et al., 2005). The importance of software estimation is centred on four dimensions: cost estimation, effort estimation, size estimation and schedule estimation (Nasir, 2006).

3.4.4.1 Data collection

We initially constructed a dataset consisting of 10 PDF files to run the extractor and assess the validity of the extractor results, which was later expanded to conduct the scalability experiment (Section 6.3). The following keywords were used to locate the articles: “Exome sequencing experiments”. The total number of studies/results obtained in this process was 33597. Several inclusion criteria were applied: all files had to be based on wet-lab and exome sequencing experiments, and all files had to have been published within the last ten years. During the data collection process, our main focus was on the title and abstract of each article.

After constructing and checking the validity of the extractor output, we expanded this dataset to be 1024 PDF files to conduct scalability experiment. We followed the same procedures as described above to select the files. All files were downloaded and stored on a local device since the experiment was conducted locally.

3.4.4.2 Performance Evaluation

Determining the scaling strategy and metrics can assist in evaluating several scalability dimensions, including cost, capacity, and quality. The scalability metric reflects the system’s ability to maintain productivity, meaning that if a system maintains its productivity as the scale changes, it can be considered scalable (Jogalekar and Woodside, 2000). Response time and throughput are key metrics in the performance evaluation of systems (Williams and Smith, 2005). Additionally, several studies have aimed to develop models for measuring and comparing costs, capacity, elasticity and other factors (Al-Said Ahmad and Andras, 2019).

In this experiment, we aimed to scale the extractor to include all articles on exome sequencing experiments in PubMed. Scalability was assessed by evaluating the performance of the extractor

as the dataset size (number of files) increased. To evaluate scalability, several experiments were conducted to measure two performance metrics: cost and response time, based on dataset size, as shown in Figure 6.6. These experiments allowed us to estimate the cost and response time required for processing each file as well as the error rate. The primary focus of this evaluation was to explore the relationship between the number of files, cost and response time.

3.5 Methodology for a User Experience Evaluation of Biomedical Dataset Search Enhancement Using Provenance Information (RQ4)

To address this final research question, semi-structured interviews were chosen to conduct a human evaluation of the provenance extractor that outlined in Section 3.4. The importance of expert evaluation has emerged as an important part in assessing the content of tools (Berk, 1990). User experience evaluation can capture users' opinions about a product, covering several aspects such as experiential and affective aspects (Vermeeren et al., 2010). Here, the term "product" refers to systems, products, and services (ISO/TC 159/SC 4, 2010). The qualitative research approach was again employed through user evaluation. Since the extractor is specifically designed for the biomedical research domain, the target population consists of biomedical researchers, whose expertise is essential for evaluating the extractor's output. We selected to conduct semi-structured interviews as this approach provides an in-depth and holistic understanding (Creswell and Creswell, 2017). By using this research method, we provided our participants with the opportunity to extract provenance information from their scientific publications and to evaluate the accuracy and correctness of this information.

The participants interviewed in this study were recruited based on their responses to our earlier survey (see Section 3.3). In addition to its primary purpose — assessing the effectiveness of providing provenance information — the survey also served as a recruitment tool, inviting respondents to provide their email addresses at the end of the survey. The inclusion criteria were as follows: biomedical researchers holding roles such as professor, research fellow, or postdoctoral researcher, and researchers involved in searching for and reusing datasets. A brief overview of participant demographics is presented in Table 7.1. The participants demonstrated a high level of expertise across various biomedical fields and had conducted multiple experiments resulting in published findings and datasets.

3.5.1 Interview design and development

In the interviews, we omitted additional questions regarding demographic information, as these details had already been collected through the survey. As mentioned in Section 3.3, the second part of the survey asked participants to provide demographic details, including their research areas and the seniority of their roles. The interview was structured into two main parts: assessing the extractor and gathering user feedback. Therefore, we began by asking the participants to provide three papers either from their own work or related to their research. These papers were uploaded to the extractor, which then provided the provenance information. Figure 3.11 illustrates the interaction workflow between users and the extractor.

The provenance information for each paper was printed, and participants were provided with different coloured highlighters, as shown in Figure 3.10. The main guideline instructions were as follows:

- Highlight all correct provenance information for each paper using YELLOW.
- Highlight all incorrect provenance information using ORANGE.
- Highlight and write any missing provenance information using BLUE.
- Highlight any repeated provenance information using PINK.

After completing the first part of the task, which involved assessing the outcome of the extractor, participants were asked to provide feedback on the extractor based on four main aspects: correctness, completeness, relevance, and reasonableness. These terms were previously defined to assess provenance quality as follows:

- Correctness: “the dimension of provenance quality denotes the extent to which provenance is correct and free of error. This dimension encompasses attributes such as the accuracy, unambiguity, consistency, and homogeneity of the provenance” (Cheah and Plale, 2014, P.4).
- Completeness: “We denote completeness as the extent to which provenance is missing or to which provenance is more than the actual amount of collectible provenance or “overcomplete” ” (Cheah and Plale, 2014, P.4).
- Relevancy: “Relevancy is the extent to which provenance is relevant and helpful to consumer needs” (Cheah and Plale, 2014, P.4).

Entities (E):

1. Pluripotent mouse embryonic stem cells (E1) - Description: Cells cultivated in specified culture mediums. - Obtained from: Wild-type male embryonic stem cell line v6.5 and Nanog reporter cell line NHET.
2. 0i Culture Medium (E2) - Description: Dulbecco's Modified Eagle Medium supplemented with specific components for cell growth. - Usage: Used for cultivating pluripotent stem cells.
3. 2i Culture Medium (E3) - Description: 0i medium supplemented with additional small molecule inhibitors. - Usage: Used for maintaining \"ground state\" pluripotency in stem cells.
4. Reporter Genes (E4) - Description: Genetic constructs designed to express a fluorescent protein in response to Nanog expression. - Types: Includes GFP-IRES-puro constructs.
5. Antibodies and Fluorescence Labels (E5) - Description: Various antibodies and dyes used for immunostaining and flow cytometry to quantify protein levels.

Activities (A):

1. Cell Cultivation (A1) - Description: Process of growing pluripotent mouse embryonic stem cells in specific culture conditions (0i and 2i mediums). - Conducted at: The laboratories involved in the study.
2. Nanog Expression Measurement (A2) - Description: Assessment of Nanog protein levels using immunostaining and flow cytometry. - Objectives: To compare expression patterns in wild-type versus reporter cell lines.
3. Genetic Reporter Design and Validation (A3) - Description: Creating and validating reporter constructs to study Nanog expression dynamics. - Involvement: Use of GFP-IRES-puro constructs in knock-in reporter systems.
4. Mathematical Modeling (A4) - Description: Development of mathematical models to understand the implications of reporter systems on Nanog dynamics. - Purpose: To explain variability in expression patterns and predict systematic perturbations.
5. Differentiation Experiments (A5) - Description: Conducted to observe changes in Nanog and GFP levels after LIF withdrawal over time. - Aim: To assess how differentiation affects the accuracy of reporter outputs.

Agents (AG):

1. Research Team (AG1) - Description: Group of researchers who conducted the study on Nanog fluctuations. - Members: [REDACTED]
2. Cell Lines Providers (AG2) - Description: Entities supplying cell lines for the experiment. - Example: [REDACTED]

PROV-DM Relationships: 1

- Prov:wasAssociatedWith (A1, AG1) - Cell Cultivation was conducted by the Research Team.

FIGURE 3.10: Example of extractor evaluation by a participant.

- Reasonability: “The degree to which a data pattern meets expectations” (Black and van Nederpelt, 2020, P.91).

3.5.2 Interview procedures

As mentioned earlier, the participants for this study were recruited using the survey. Initially, we obtained interest from nineteen biomedical researchers who wished to participate. All those researchers were contacted via email to schedule interviews, but we received responses from only ten of them. This number still falls within the recommended range of 5 to 25 participants (Lilly, 1998; Dudovskiy, 2016). Brief demographic information is provided in Table 7.1, reflecting the participants' fields and roles. All interviews were conducted between September 2024 and

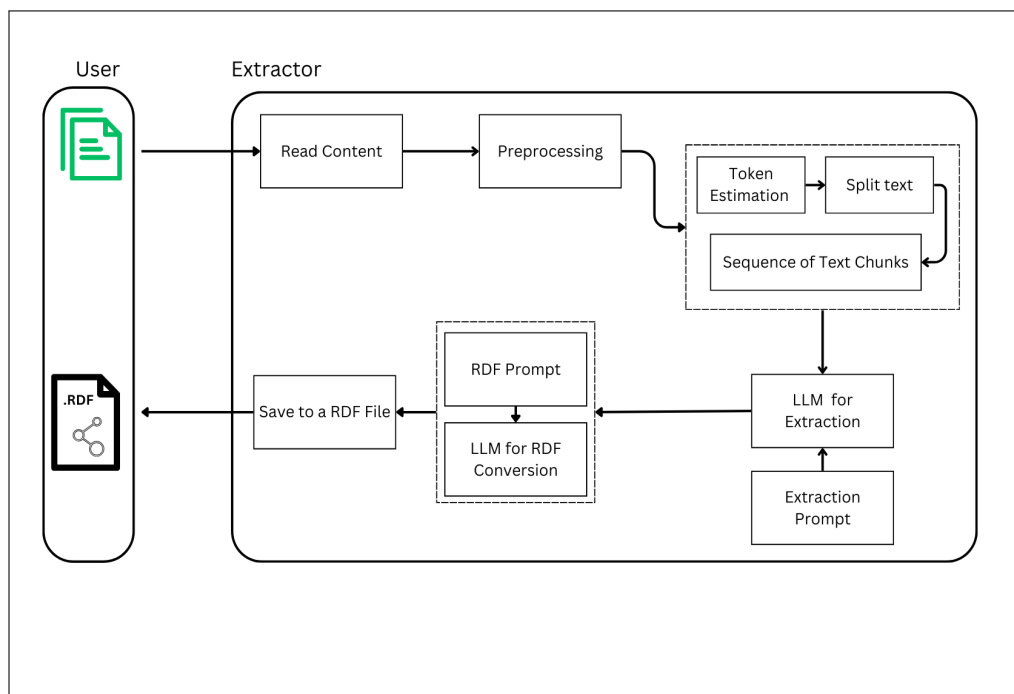


FIGURE 3.11: Interaction workflow between users and the extractor.

October 2024. Each interview lasted approximately 45 minutes, and all were conducted in English. The personal information was anonymized to ensure participant privacy.

We interviewed ten biomedical researchers, each of whom provided three papers for evaluating the extractor. All the papers fall within the field of biomedicine and are based on wet-lab “Omics” experiments. To ensure participant privacy, all publications used with the extractor tool were anonymised before analysis. Only the textual and structural content necessary for provenance extraction was used.

3.5.3 Interview Analysis

Since the aim of the interviews were to evaluate the extractor and explore how the participants feel about the results, a mixed-methods approach was used on the interview results. First, we conducted quantitative analysis to evaluate the extractor outputs. Second, a thematic analysis was used as well to organise the results and the feedback, as it is one of the most common strategies for qualitative research (Robson and McCartan, 2016).

For the quantitative analysis in this study, the collected provenance information extracted during the interviews was modeled using the PROV-DM format: Entities, Activities, Agents and Relationships. First, this provenance information components (i.e. Entities, Activities, Agents and Relationships) categorised into four themes: correct item, incorrect item, missing information

item and repeated item. Secondly, we calculated the accuracy of each extracted provenance information using classic information retrieval metrics (**Precision** and **Recall**).

$$\begin{aligned} Precision_{Comp} &= \frac{|P \cap Pe|}{|P|} \\ &= \frac{\text{Proportion of relevant provenance information retrieved from the extractor}}{\text{All provenance information retrieved from the extractor}} \end{aligned} \quad (3.3)$$

where the provenance information can be entities, activities and agents. There is provenance information returned from the LLM (p) which contains {e,a,c,r} and provenance from an expert (Pe) which contains {ee, aa, ce, re}.

$$\begin{aligned} Recall_{Comp} &= \frac{|P \cap Pe|}{|Pe|} \\ &= \frac{\text{Proportion of relevant provenance information retrieved from the extractor}}{\text{All relevant provenance information provided by the evaluator}} \end{aligned} \quad (3.4)$$

We observed in some interviews that the components were correct, but their descriptions were either incorrect, partially missing, or repeated. Therefore, we measured the provenance information twice: the first time, we calculated the provenance information components (e.g., Entities) as correct items even if their descriptions were incorrect, partially missing or repeated; and the second time, we took the correctness of the descriptions into account. Thus, we designed additional metrics:

$$\begin{aligned} Precision_{Desc} &= \frac{|P \cap Pe|}{|P|} \\ &= \frac{\text{Proportion of relevant description for each provenance component retrieved by the extractor}}{\text{All provenance component descriptions retrieved by the extractor}} \end{aligned} \quad (3.5)$$

$$\begin{aligned} Recall_{Desc} &= \frac{|P \cap Pe|}{|Pe|} \\ &= \frac{\text{Proportion of relevant description for each provenance component retrieved by the extractor}}{\text{All provenance component descriptions provided by the evaluator}} \end{aligned} \quad (3.6)$$

To conduct thematic analysis, [Braun and Clarke \(2006\)](#) recommend following six steps, which is explained in Section 3.2.3. The advantage of using thematic analysis lies in its flexibility, which allows researchers to gain a rich, detailed understanding of the results. This analysis was conducted using NVivo12. As mentioned earlier, there are two coding methods: inductive and deductive approaches. In this analysis, we followed an inductive approach to explore the experts' opinions about the completeness, correctness, relevancy and reasonability, and collect their suggestions. The coding process resulted in several main themes and sub-themes.

3.6 Ethical Approval

All participant-based studies in this research obtained ethical approval. The approval was granted by the University of Southampton Ethics Committee for the following three studies:

- **Dataset Search for Biomedical Researchers - (ERGO/FEPS/73032)**

Since all interviews were conducted online, the ethical approval documents, consent form and participant information sheet, were sent in advance for participants to read and sign. The consent form included several statements: confirming that the information sheet was read and understood, agreeing to participate in this research project, consenting to the recording of the interview and acknowledging that participation is voluntary with the option to withdraw at any time.

- **A User Study on the Effect of Provenance Information Availability on Biomedical Dataset Search - (ERGO/FEPS/92985)**

The consent form and a link to the participant information sheet were integrated into the survey. Before participating, all respondents had to confirm that they had read and understood the information sheet. Only after providing this agreement could they access the survey.

- **A User Experience Evaluation of Biomedical Dataset Search Enhancement Using Provenance Information - (ERGO/FEPS/98745)**

All interviews were conducted in person; therefore, ethical approval documents were provided earlier for participants to read, understand and sign prior to the evaluation. All statements mentioned above were also included in this form.

3.7 Summary

In this chapter, we discussed the research methodologies used to answer each research question. First, we presented the method for identifying the dataset search requirements of biomedical researchers. Second, we provided the methods aimed at addressing the requirements of biomedical dataset search. We also outlined the methodology used to assess the effectiveness of incorporating provenance information in dataset search for biomedical research. Third, we discussed the methods and tools used to build the extractor, which was designed to automatically extract provenance information to meet the requirements of biomedical dataset search. Finally, we presented the methodology for evaluating the extractor's performance and scaling it.

Chapter 4

Dataset Search for Biomedical Researchers

In the following section, we explain the key findings that emerged from the data collection and analysis detailed in 3.2.1. Our analysis has identified six main themes for understanding dataset search and usage in biomedical research: the variety of data formats and types; the importance of metadata; issues pertaining to data quality; dataset accessibility; the methodologies employed in the assessment of datasets; and the strategies used for dataset search. Following the prevalence of provenance in our interview data, we devote special attention to discussing provenance needs. We conclude this chapter by presenting the query styles used among researchers in their quest for appropriate datasets.

4.1 Demographics

We initially interviewed 17 biomedical researchers, however, three participants were excluded from the study. Two of the three were project assistants who had been provided a dataset by their line managers and were thus not involved in the search process, and the third was a new PhD student who had not yet gained experience in locating datasets. Another interviewee was excluded as they withdrew and apologised for not being able to complete the interview. As stated earlier in Section 3.2, we aimed to understand how biomedical researchers currently search for datasets, the challenges they face with these methods, and their requirements for improving dataset search.

Participants	Research Domains	Roles
P1	Genomics	Research Fellow
P2	Human Development and Health	Postgraduate researcher
P3	Clinical and Experimental Science	Professor
P4	Biomolecular Medicine	Professor
P5	Genomics	Professor
P6	Genomics	Senior Research Fellow
P7	Human Development and Health	Postgraduate researcher
P8	Human Development and Health	Professor
P9	Clinical and Experimental Science	Postgraduate researcher
P10	Cancer Sciences	Lecturer
P11	Human Development and Health	Postgraduate researcher
P12	Cancer Science	Postgraduate researcher
P13	Cancer Science	Senior Research Fellow

TABLE 4.1: Demographic information of participants in study 1

The total number of acceptable interviewees was thirteen biomedical researchers. Four of whom were professors working in different domains, including Genomics and Biomolecular Medicine. Five of the interviewees were senior researchers engaged in biomedical data-driven research, whilst the remaining four were advanced postgraduate researchers in their third or final year of study. These postgraduate participants were supervised by multiple professors. All participants were members of the Faculty of Medicine at the University of Southampton. Among the participants, 70% were senior researchers holding various roles, including professors and research fellows, while the remaining 30% were postgraduate researchers.

The interviewees were experts in various research domains, including clinical experimental science, cancer science, genomics, bowel diseases, microbiology, and biomolecular medicine. Table 4.1 presents the research domains and roles of all participants, as reported during the interviews. After conducting 13 interviews, we reached the saturation point, as no new information emerged. This is consistent with what has been reported in the literature, as illustrated in Section 3.2.

4.2 Findings

4.2.1 Data Formats and Types

All participants (P1–P13) reported that their search processes and requirements often focused on highly structured datasets. One participant (P12) noted that some researchers within their research group faced several challenges when working with other types of data, including

semi-structured or unstructured data; therefore, structured data were generally preferred. For this reason, they tended to search for more structured data within domain-specific repositories, such as DataMed, which typically include these types of datasets. One participant (P9) stated that the use of structured data helped facilitate the analysis process. Only one participant (P10) stated that their needs sometimes required the use of a combination of structured and unstructured data to obtain more comprehensive information, including handwritten notes shared between doctors in PDF format, which are classified as unstructured data.

As all interviewees in this study were biomedical researchers (P1–P13), they often search for sequencing data necessary to support their tasks. Sequencing data is a type of biological data that can present various types of information, including DNA, RNA, and proteins, which can be used to identify variations that may cause disease (Quinn et al., 2018). For example, participants (P5, P7, P12) reported searching for genome sequencing datasets, while participants (P12, P13) focused on transcriptome data. Additional participants (P1, P2, P3, P6, P8, P10, P11) reported searching for DNA, RNA, and exome sequencing datasets.

As a result, all participants expressed the need to discover more detailed information about the sequencing data within datasets prior to accessing them. They emphasised the importance of gaining a deeper understanding of the data collection process. Additionally, they highlighted the need to explore the background of sequencing pipelines — that is, the steps followed to generate the sequences from raw data. Therefore, this requirement supports researchers in understanding the content of datasets and making informed decisions. This claim aligns with Stamatogiannakis et al. (2017), which states that provenance information can improve decision-making processes. Additionally, it can improve the comprehensiveness of metadata within the research community, thereby facilitating the dataset search process. Furthermore, it underscores the importance of incorporating provenance information into metadata, referred to as provenance metadata.

Despite multiple efforts to standardise data formatting in various biomedical research communities that utilise and manipulate datasets, researchers continued to encounter issues when dealing with different data formats. One participant (P7) stated that this lack of consensus on formats results in significant time being wasted during preprocessing. This aligns with prior research, which highlights that biological researchers face challenges with complex and varied data formats (Anderson et al., 2007). This issue can lead to delays and inefficiencies. A participant (P9) expressed their colleagues' concerns about the time wasted in handling various formats. Another interviewee (P6) mentioned their preference for formats that are easily compatible with their existing tools. Furthermore, several researchers reported attempting to integrate data from diverse sources and formats to construct suitable datasets tailored to their specific tasks, frequently encountering challenges related to data integration (P7, P9).

Several common data formats and standards, including GTF, FASTQ, and BED File Format, are used in biomedical fields for genomic data. In addition, several participants (P1, P4, P10) stated that they prefer to utilise common dataset formats, including CSV, Excel, and TXT. However, one primary reason for the proliferation of various dataset formats is the lack of consensus within the community. System developers and data suppliers frequently do not adhere to the data or dataset standards prevalent in their communities (P10).

4.2.2 Query Styles

Our study reveals that all participants predominantly utilise keyword search as a query style within the search engines, data portals, and repositories they access. Query style refers to the form of input provided to the search engine by the data seeker, as described in the taxonomy by [Almuntashiri et al. \(2022\)](#). This method is used whether searching for a dataset or publications related to the datasets. It is the most immediate way to interact with repositories and is often the only option when using public sources.

After initiating a keyword search, several participants in our study reported adopting more sophisticated search techniques. One widely used method is filtering, a feature commonly available in most search engines and publicly accessible portals (P11, P6, P4, P3, P7). However, there is no standardisation of the facets implemented by each portal, except for publication year and geographic areas (where applicable).

4.2.3 Strategies and Places

The participants were asked to discuss the strategies they use to search for datasets based on their needs. Additionally, they were asked to identify any differences among public sources, systems, and websites. We identified three main strategies used among all participants, as presented in Figure 4.1. These strategies are used sequentially to search for desired datasets.

The first strategy involves searching for datasets through public sources, including common search engines such as Google Dataset Search¹, open data portals such as GEO², and public repositories such as GenBank³. Based on the required data type, researchers search for specific repositories and portals; for instance, MetaboLights⁴ is a public repository dedicated to metabolomics

¹<https://datasetsearch.research.google.com/>

²<https://www.ncbi.nlm.nih.gov/geo/>

³<https://www.ncbi.nlm.nih.gov/genbank/>

⁴<https://www.ebi.ac.uk/metabolights/>

experiments. In this study, one interviewee (P1) searched the GTEx Portal⁵ to retrieve datasets that include RNA sequencing data for different tissues, while another participant (P11) searched for exome datasets within UK Biobank⁶. However, biomedical researchers may encounter difficulties in generating queries that characterise their needs in this strategy (Dixit et al., 2018).

The second strategy involves reviewing literature, journals, and publications. The aim of this approach is to search for new data or datasets (P5, P9, P13, P12, P1), to gather more details about recent datasets (P6, P7, P3, P12, P8, P10, P1), or to find datasets and studies addressing similar topics (P2, P13). Dataset search are often associated with literature searches to investigate related publications in order to gather additional information that can assist in assessing relevance and quality (Krämer et al., 2021).

The final strategy used to search for datasets includes social communication. This means that researchers contact dataset owners to request access or share their datasets. Such communication is typically conducted through emails, discussions at meetings, or academic events. This strategy is not only employed to search for datasets but also to obtain additional information about them, including provenance information.

Our observations indicate that most participants followed these three strategies in a consistent order when conducting the dataset search, as illustrated in Figure 4.1. Participants typically initiate their search through open portals or public repositories specific to their research domain, either by navigating to well-known sources in their community or by using search engines — such as DataMed — to search for relevant datasets. This strategy can save time by reducing the need to read publications or conduct experiments to generate datasets. Subsequently, they move to the second strategy: consulting literature and publications to find datasets, identify similar studies, or obtain more information about a particular dataset. This strategy can complement the first strategy (dataset search) to find additional information if the metadata is incomplete. Finally, participants seek assistance from field experts or dataset holders to share datasets or conduct further investigations. This strategy can help researchers obtain specific descriptions or explanations about retrieved datasets.

4.2.4 The Importance of Metadata

An important theme consistently emphasised across all interviews was the significance role of metadata in dataset search. Several participants in this study (P1, P4, P5, P6) confirmed their use of metadata during the dataset search process. The use of metadata helps researchers assess

⁵<https://www.gtexportal.org/home/>

⁶<https://www.ukbiobank.ac.uk/>

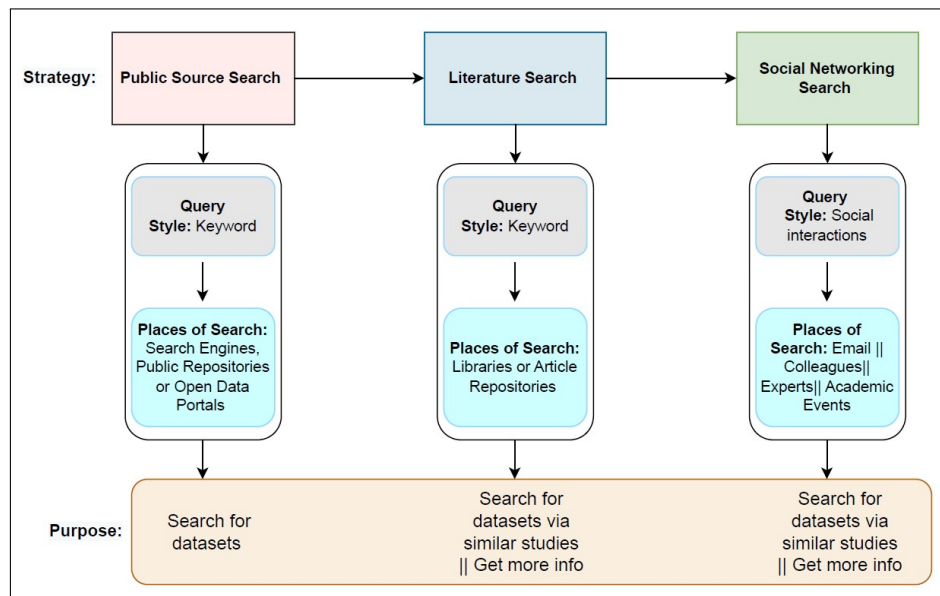


FIGURE 4.1: Order of search strategies followed by our participants

Metadata

Name
Familial Exome Sequencing in Rare Pediatric Phenotypes

Repository
dbGaP

Identifier ←
Unknown

Description
To discover novel candidate genes associated with rare Mendelian phenotypes, we will conduct individual genomic and phenotypic characterization using genome-wide array, pedigree exome sequencing, candidate genotyping, and pertinent clinical testing to define phenotype. Pedigrees included in this submission will have a variety of clinical pathological phenotypes.

Data or Study Types ←

Source Organization
National Heart, Lung, and Blood Institute DAC

FIGURE 4.2: Example of missing metadata elements

a list of datasets in order to identify the most suitable ones based on their specific needs (P1, P2, P5, P6, P8, P12). However, several participants (P3, P8, P12) noted a lack of sufficient detail or critical information in the current metadata. This finding aligns with [Dixit et al. \(2018\)](#), who emphasises that existing metadata often fails to adequately describe datasets, as it typically includes only a part of crucial information. Figure 4.2 presents an example of missing metadata extracted from an open portal.

Many participants highlighted their requirements for which data should be included in metadata. A notable prerequisite highlighted by many participants is data provenance. Desired provenance information includes the criteria for the inclusion/exclusion of samples in the dataset (P9), its origin (P1), the identification and handling of these data (P9), how the evaluation phase was conducted (P11, P8), and its size (P1, P2, P9). In light of the importance our interviewees placed on provenance, we conducted a deeper analysis in Section 4.2.8, titled ‘Expansion of Provenance Concerns’.

Despite many efforts, the standardisation of metadata remains a common issue in this community. We observed that the level of standardisation depends on the type of data; for instance, standards are more advanced for imaging data than for genomic data (P5). This issue can arise due to various factors, including the diversity of data types, the requirements of each domain, methods of data generation, analysis or sequencing techniques, and the presentation of metadata in portals. Some participants (P9, P13) confirmed that a universally standardised metadata format is still needed to accommodate all types of biomedical data.

4.2.5 Data Quality Issues

Data quality concerns emerged as a critical challenge during dataset search, as reported by several participants (P4, P8, P9, P10). One participant (P10) stated that researchers typically search for high-quality datasets, as these facilitate their use without the need to perform a pre-processing phase. This can prevent the need to collect data from scratch, saving time, effort, and resources. Additionally, several participants mentioned that while they may initially find datasets that meet their requirements, quality issues often arise upon further investigation, leading them to avoid using these datasets (P6). Researchers find the process of assessing data quality difficult and time-consuming (P5, P6). Consequently, they sometimes tend to trust the quality of datasets without fully verifying them, which can lead to producing incorrect results or inconsistent results.

Data preparation is a significant challenge due to data quality concerns, which in turn makes it difficult to utilise the retrieved datasets (P1, P7, P10, P4, P12, P2). Other participants emphasised how time-consuming data preparation is (P10, P7, P5). One participant (P1) confirmed that dataset cleaning presents a challenge due to the significant amount of noise in the data. Several factors such as language issues, data incompleteness and other factors may contribute to this challenge. Two participants (P1, P5) expressed interest in being informed about the measures of noise in the data and any pre-processing that has already been conducted, to better focus their own pre-processing efforts. However, data preparation is essential step for yielding high-quality data (Zhang et al., 2003).

Data quality concerns varied, depending on the context and nature of each research project. Therefore, it was difficult to identify a specific issue in this study. We observed two common issues during dataset search: data completeness and variations in units of measurement.

The first concern, **data completeness**, is defined in the Data Management community (Black and van Nederpelt, 2020) as the extent to which dataset records contain all essential and expected data needed. Data completeness is an important aspect as it significantly affects the accuracy of outcome (P3, P5, P86, P8, P13). Furthermore, missing fields or records were identified in this study as a common problem, making the data harder to use (P9) and influencing selection (P5). Figures 4.3 present an example of incomplete data retrieved from a public biomedical repository.

MMWR	Year	MMWR	Week	Chlamydia	Chlamydia	Chlamydia	Chlamydia	Chlamydia	Chlamydia	Chlamydia	Chlamydia	Chlamydia	Chlamydia	Coccidioidomycosis	Coccidioidomycosis	Coccidioidomycosis	Coccidioidomycosis	Coccidioidomycosis	Coccidioidomycosis	Coccidioidomycosis	Coccidioidomycosis
2016	1	11455		28656		31128		11455		25886		139		208		353		139		176	
2016	1	360		935		1429		360		857		-		0		0		-		-	
2016	1	1735		3589		4220		1735		3095		-		0		0		-		-	
2016	1	266		1193		1353		266		1295		N		0		0		N		N	
2016	1	1664		4193		4587		1664		4164		-		1		2		-		-	
2016	1	343		1686		1934		343		1748		-		2		7		-		-	
2016	1	1360		6039		7647		1360		4754		-		0		1		-		-	
2016	1	-		132		403		-		133		-		0		0		-		-	
2016	1	161		1472		2600		161		1107		-		0		0		-		-	
2016	1	2897		3720		6545		2897		3481		-		0		2		-		-	
2016	1	567		1844		2514		567		1798		139		149		284		139		89	
2016	1	2368		4694		5208		2368		4882		-		63		94		-		87	
2016	1	-		-		-		-		-		N		-		-		N		N	
2016	1	-		-		-		-		-		-		-		-		-		-	
2016	1	-		15		26		-		12		-		0		0		-		-	
2016	1	20		82		123		20		120		N		0		0		N		N	
2016	1	115		317		434		115		314		-		0		1		-		1	
2016	1	-		284		350		-		314		N		0		0		N		N	
2016	1	174		562		693		174		493		-		0		1		-		-	
2016	1	10		66		103		10		73		N		0		0		N		N	
2016	1	721		426		3351		721		138		-		0		1		-		-	
2016	1	161		578		794		161		582		N		0		0		N		N	
2016	1	434		558		938		434		444		N		0		0		N		N	
2016	1	7		325		686		7		3		-		0		1		-		-	
2016	1	68		142		267		68		267		-		0		1		-		-	
2016	1	-		1041		2547		-		1460		N		0		0		N		N	
2016	1	1009		1029		1465		1009		860		N		0		0		N		N	
2016	1	54		108		154		54		87		N		0		0		N		N	

FIGURE 4.3: Example of an incomplete dataset from NNDSS (2016), with missing information in several columns, including data on Chlamydia trachomatis infection and Coccidioidomycosis.

Additionally, one participant (P3) noted a lack of association between publications based on data within dataset and the uploaded datasets themselves. Liu et al. (2016) stated that data incompleteness can occur when data is repurposed from commercial to scientific use. In addition, several restrictions can affect data completeness, such as availability or publishing restrictions.

The second common quality issue identified in this study is **the variance in units of measurement** within datasets, as noted by several participants (P5, P9, P10). This concern can negatively affect the usability of datasets found in a search process, primarily due to the extra time required to standardise units across various datasets before beginning any analysis. One participant (P10) stated that, in the healthcare domain, there is an absence of standardised measurement units

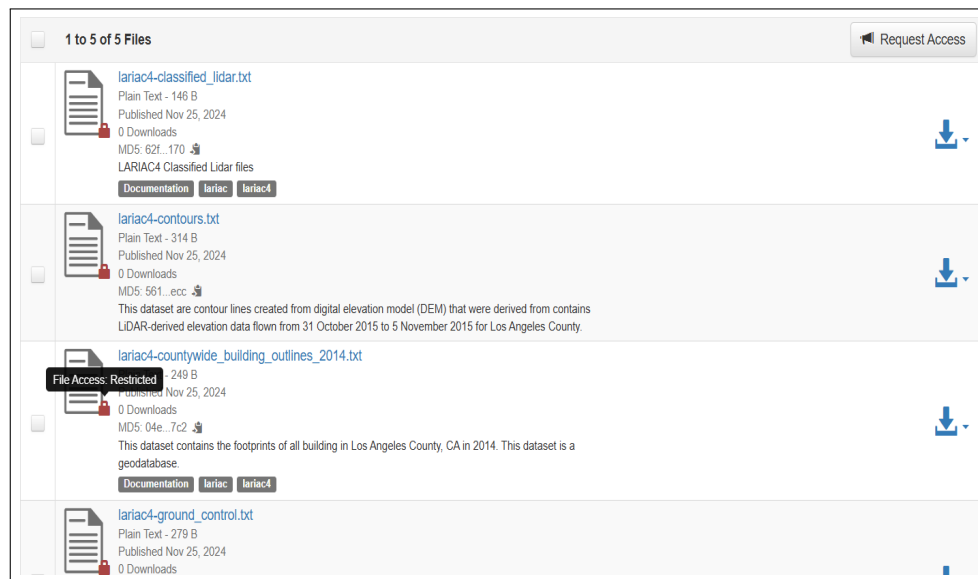


FIGURE 4.4: Example of access restrictions from an open data portal

applied to the data collected, which can hinder data integration and comparison. Another participant (P5) pointed out that the main challenge is the diversity in data collection methods of different laboratories or hospitals. Additionally, two participants (P5, P9) mentioned their need to determine if any data is missing in advance, in order to update the dataset consistently and to understand the methods used to measure the existing data.

4.2.6 Accessibility

A new crucial challenge influencing the dataset search process is accessibility. This concept is defined as the ease with which data can be accessed and retrieved (Wang and Strong, 1996). Ding et al. (2014) confirmed that there is a gap between users' needs and dataset accessibility. The significance of accessibility during dataset search is mentioned as a concern by the majority of our participants (11 out of 13). They noted that accessibility can enhance the process of searching for datasets online.

One prevalent issue affecting dataset search is the lack of clear metadata on how to access datasets. Often, accessing datasets requires users to comply with various policies and regulations, as well as obtain approval from the data owner. Figure 4.4 presents an example of access restrictions to datasets available on an open platform. Unfortunately, details about these policies, regulations, and the steps for obtaining approval are not always easily accessible (P1). Xiao et al. (2019) mentions in their questionnaire that dataset searchers look for detailed documentation related to datasets within data portals. They need help understanding the datasets they want to engage with.

In addition, one participant (P6) noted that the process of accessing a dataset can sometimes be more challenging than dealing with the dataset being unavailable or inaccessible. This difficulty can stem from the design of the portals or repositories themselves. In addition, it can also stem from the link between search engines and dataset repositories. Two participants (P2, P12) mentioned that the ease or difficulty of accessing datasets can vary depending on the portal. Our findings here is compatible with challenges mention in (Xiao et al., 2019). In addition, Zuiderwijk et al. (2012) indicated that users face challenges in understanding open government data due to a lack of information needed to interpret it.

Dataset accessibility is closely linked to the ability to download the files, which is essential for researchers to perform their tasks. Many participants reported difficulties in downloading datasets (P1, P10, P13, P7). Several challenges were mentioned, including broken links (P7, P11) and incorrect uploads by data owners (P6). Due to concerns around user privacy and confidentiality, facilitating access to datasets for scientific purposes can be restricted. Consequently, several organisations require users to obtain authorised access to their datasets, owing to data privacy considerations. As a result, such datasets are often only accessible within controlled or individual research environments. Researchers seeking access often need explicit legal approval (P11). For example, two interviewees (P9, P10) experienced difficulties accessing datasets held by a university. Access to these datasets required the submission of an ethics application. The process of obtaining this approval can be lengthy and cause delays, posing significant challenges, as noted by one participant (P13).

4.2.7 Dataset Assessment

A critical requirement was identified as a major concern by our participants across their research groups. The majority of our participants (12 out of 13) expressed a need to enhance this aspect in order to facilitate their dataset search. Three participants (P1, P2, P6) highlighted the obstacles related to evaluating the suitability of datasets. In other words, they experienced difficulties in determining which of the retrieved datasets were most appropriate for their research needs. Other participants (P5, P6, P8) also highlighted the absence of formal methods for evaluating datasets.

Several current judgment techniques and methods for assessing or evaluating retrieved datasets are introduced by the majority of participants (9 out of 13). The first technique is **the metadata of datasets** during the assessment phase, as it assists in judging and reusing the datasets. However, metadata incompleteness remains an obstacle for dataset seekers. Two participants (P5, P6) noted that incomplete metadata hindered their ability to accurately assess datasets, underscoring the importance of metadata in dataset search.

Another widely used technique involves **searching for and reviewing literature**. More than half of the participants (P1, P8, P4, P5, P7, P13, P11) reported thoroughly examining literature to aid in dataset assessment. This technique can be divided into two approaches: searching for published papers associated with the dataset to obtain necessary information and reviewing similar works or publications to inform their judgment.

Domain-specific experience is another strategy for judging datasets. Additionally, consulting experts in the field is a common technique to aid in the assessment process. Two participants (P5, P10), who relied on their extensive experience (over 20 years), emphasised this approach. Another participant (P7), with less experience, sought guidance from more senior colleagues to inform their judgment.

Finally, several interviewees reported using a technique, **manual examination of datasets**, to evaluate their comprehensiveness (P3, P6), completeness (P9), size (P9), and to compare them with other datasets (P2). Additionally, with the popularity of **ML techniques** across most domains, our participants in this study reported using such techniques to evaluate and assess the datasets they retrieved. Examples include hierarchical clustering (P12) and data visualisation (P1, P7) to evaluate the quality of the data contained within the datasets.

Several methods for assessing datasets have been proposed in other fields, such as a technique designed to evaluate the coverage of a given dataset to identify vulnerabilities (Asudeh et al., 2019), and a framework for assessing observational climate datasets (Zumwald et al., 2020). Although the ways to assess datasets were mentioned by our participants, there is a gap in automated frameworks and techniques that needs to be addressed in our domain.

4.2.8 Expanding the Importance of Provenance for Biomedical Researchers

An important aspect identified as a requirement by participants in this study is the importance of providing provenance information. All participants (P1–P13) reported the need to understand the history of the datasets they retrieved in order to determine whether those datasets are relevant to their tasks. Additionally, participants emphasised that providing provenance information is a critical requirement in dataset search domain. They pointed out that provenance can help researchers understand the strengths and weaknesses of datasets.

Several interviewees (P1, P11, P10, P3, P6, P12) explicitly emphasised the clear need for provenance, while others provided specific reasons for its importance. Some aimed to use provenance to discern the differences between retrieved datasets and those already owned by researchers (P9). Provenance information can be used to gain more knowledge about the techniques, tools,

procedures utilised in the lab when collecting the data (P11, P1). For example, one participant (P10) sought provenance information specifically because they relied on the data for patient treatment, using it to understand the history of the datasets. These tasks necessitate the use of multiple datasets or those of a particular size, thereby requiring the integration of various datasets. Providing provenance alongside datasets can facilitate comparison between them, enabling researchers to integrate the datasets more effectively for their tasks (P2, P6, P8).

Although significant efforts have been made in provenance development over the past decade, the absence of provenance information during dataset search was highlighted by (8 out of 13) participants in this study. During this study, we discovered a lack of meeting users' needs for provenance information in this field. This includes exploring datasets or describing the experiments on which the datasets are based.

Several aspects, as mentioned by participants, can be affected by the lack of provenance: **dataset evaluation**, often leading to its exclusion (P3, P8); **processing time** (P13, P7); and **the analytical and reasoning processes of research** (P1).

The lack of provenance information is consistent with the literature, which mentions that there is a gap in the biomedical research domain that needs to be addressed ([Buneman and Tan, 2007](#); [Collins and Tabak, 2014](#); [Baum et al., 2017](#); [Liu et al., 2020](#)).

The participants outlined the requirements for provenance, which can vary depending on users' needs. However, all requirements fall within the scope of provenance, such as the methods used to generate the data. As part of this study, we asked participants to identify the types of provenance information that could help enhance the dataset search domain. Table 4.2 outlines the stated requirements for provenance information.

- The first provenance requirement is **how the data within the dataset were generated and built**. This includes the methods used for data collection or generation within the dataset and how the dataset were constructed.
- Another requirement is access to more information about **the techniques used for processing the data within datasets**. Several participants noted that this includes the technical methods employed, the types of samples used, and the analytical procedures followed. Participants (P1, P5, P11, and P13) emphasised the need to acquire more detailed information about the technical tools and equipments used during the experiment to process and generate the data within the datasets.
- Moreover, several participants expressed a desire for information about **when and where the data or dataset comes from**. “When” and “where” requirements are common needs

Requirements for Provenance Information	Participants	Example quote
How is the dataset built or collected?	P1, P8, P10, P2, P4, P3, P12, P9, P11, P6, P5, P7, P13	“Information on the kind of methods used to generate that data”
How is the dataset/data processed or what methods or technique were used?	P1, P8, P10, P2, P12, P11, P6, P5, P13	..“how the data was processed, includes all of the stuff in the lab..”
Where does the data/dataset come from?	P1, P8, P10, P2, P9, P7, P13	“we need to know where the data came from...”
What is the size of the dataset?	P1, P2, P9, P6, P5, P7	...“What is the size of the dataset...”
When was the data/dataset generated?	P8, P3, P12, P6	...“who entered it when they entered it...”
Who generated the dataset?	P8, P10, P4, P3, P11, P5	...“So then the source does become important because if it’s a reliable source..”

TABLE 4.2: Summary of provenance requirements elicited by our study.

in other domains, as mentioned in the literature (Johns et al., 2023; Ram and Liu, 2007; Miles et al., 2007). “When” indicates the time spent on data collection and processing. “Where” indicates information about the locations used for data collection.

- Another observed requirement is the **provision of more information about the main source of the dataset**. For example, several participants expressed a preference for knowing the source of the dataset or the identity of the dataset depositors, as this contributes to trust the dataset.
- The final requirement identified by our participants is know more information about the **size of the dataset**. This refers to knowing the actual volume of data within the dataset. Some participants noted that certain datasets are not fully uploaded and the reported size of the dataset is not always consistent with the publications based on the datasets.
- As this study focuses on the biomedical research domain, several participants expressed interest in and examined various types of data, including sequencing data. Participants (P2, P6, P11, P13) emphasised the need to understand **how the data were sequenced**, including the techniques, methods and tools used in the experiments.

Some of the above requirements for provenance information mentioned here have been previously presented in prior studies as requirements in other similar domains, as discussed in a scoping review (Johns et al., 2023). However, the authors of this review confirmed that the requirements

for provenance information can vary based on users' needs, but are similar in most use cases. Table 4.3 presents a comparison of requirements for provenance information.

Findings	Johns et al. (2023)	Our findings
Where provenance	Considered	Considered
How provenance	Considered	Considered
Who provenance	Considered	Considered
Why provenance	Considered	Not considered
When provenance	Not considered	Considered
Size of dataset	Not considered	Considered

TABLE 4.3: Comparison of provenance findings from this study with those reported by Johns et al. (2023)

To find the required provenance information, participants usually search for publications to discover and read, as publications typically include the provenance associated with the dataset. Thus, linking publications to dataset search engines or providing provenance within these engines is an important gap that needs to be addressed.

4.3 Discussion

4.3.1 Comparison to Previous Work

Gregory et al. (2020) investigated how data seekers search for and evaluate datasets, conducting their data collection in the latter half of 2017. Given the rapid rapidity of research and technological advancements, these findings may have been affected by subsequent developments. In terms of disciplines, this study encompassed participants from 17 different domains, even though each discipline has its own nature and specific requirements for assistance tools and other necessities. Additionally, it encompassed a diverse group of participants, including citizens, support staff and industrial personnel from various countries. This broad socio-technical focus contrasts with our research, which narrows its scope to biomedical researchers, particularly those specialising in fields such as genomics and cancer sciences. By targeting this specific sub-community, our study aims to identify their distinct needs and enhance dataset search functionality for biomedicine.

As shown in the comparison in Table 4.4, there are several common findings with (Gregory et al., 2020): the use of diverse search strategies, reliance on auxiliary information, various methods of accessibility, a lack of metadata quality, and a lack of evaluation techniques. These common findings confirm the importance of the dataset search domain across disciplines, specially research domains. Additionally, they emphasise that dataset seekers universally require improvements

Findings	Studies	
	Gregory et al. (2020)	This study
Use multiple search strategies	Considered	Considered
Combine search strategies sequentially	Not considered	Considered
Use auxiliary information	Considered	Considered
Missing needed data	Not considered	Considered
Utilise various methods to access datasets	Not considered	Considered
Missing dataset evaluation techniques	Considered	Considered
Need better metadata quality	Considered	Considered
Face data quality issues	Considered	Considered

TABLE 4.4: Comparison of findings from this study and prior research Gregory et al. (2020)

in current dataset search methods. These improvements can enhance the application of FAIR principles in this dataset search.

However, certain findings are unique to the biomedical research community, such as the integration of multiple search approaches for datasets, a lack of detailed provenance and significant issues with data quality. These specific insights have been translated into requirements aimed at refining the dataset search process for this particular community. A more in-depth exploration of these requirements for biomedical researchers will be presented in the following subsection.

4.3.2 Dataset Search Requirements for Biomedical Researchers

Given the significance of searching for public datasets in the biomedical domain, our participants employ a variety of search strategies to locate desired datasets or gather additional information about existing ones. We found that participants typically begin their dataset search with public sources, followed by a literature search to identify related research and determine the datasets used or to acquire additional information. Finally, they use social network search as a strategy to find datasets. These search strategies, extensively discussed in the ‘Methods and Places Used to Search for Datasets’ section, are generally employed sequentially.

Several efforts have integrated public source search with literature search to find datasets, such as the GWAS Dataset Finder (Dong et al., 2017) and the concept of dataset discovery in application contexts (Singhal and Srivastava, 2017). Additionally, other efforts have sought to combine web search with information from emails, such as the Flink system (Mika, 2005). Combining these search strategies into a single system could significantly enhance dataset searching for biomedical researchers. In terms of social network search, further developments could benefit from personalised social search based on social relations, as demonstrated by (Carmel et al., 2009). However, this technique has only been proven effective in corporate settings, where

social relations are managed by a central entity and are highly curated. In the research context, constructing an accurate social network presents a significant challenge in itself.

A significant challenge impacting dataset search for this community is the improvement of metadata quality in public sources to meet biomedical researchers' needs. Addressing this challenge can assist researchers in dataset assessment and decision-making. Consequently, they aspire to find information relevant to their needs, such as provenance information, in related publications. Such information plays an essential role in ensuring the reproducibility of research results and their trustworthiness (Gierend et al., 2024; Valdez et al., 2017).

There are several provenance models, as detailed in section 2.3.1, both generic and domain-specific, that can be used to capture, collect, identify, generate or extract data provenance. These include the W3C PROV (Missier et al., 2013a) and the W7 Model (Ram et al., 2009). For instance, McCusker and McGuinness (2010a) developed a technique for converting experimental data into a provenance representation. However, it specifically targets microarray experiments, utilising the Open Provenance Model and Proof Markup Language. Another approach aims to standardise provenance information for biological specimens and data (Wittner et al., 2020), although Wittner et al. (2024a) indicate that this standard is still ongoing. Another project, ProvCaRe (Valdez et al., 2017), extracts simple provenance records from sleep medicine papers using a standard NLP pipeline. This project, which is the most similar in style and domain, is now inoperable and specifically modelled for the sleep medicine domain.

However, there is still no established link between provenance models and the information or metadata provided by public sources. Establishing such a connection could save time by reducing the need to search for and read publications or conduct experiments to regenerate datasets.

4.4 Summary

This study has identified the current dataset search techniques used by biomedical researchers and the obstacles they encounter. Additionally, the requirements essential for enhancing dataset search in biomedical research were investigated. Table 5.4 provides a comprehensive summary of the requirements elucidated in this study. We contribute to this field by addressing one of these requirements: the advancement of metadata through the integration of provenance information in biomedical dataset search.

Chapter 5

Measuring the Effectiveness of Provenance Information for Dataset Search

In this chapter, we present the results of a user study aimed at assessing the effectiveness of provenance information in biomedical dataset search. This assessment was conducted through a questionnaire, presenting different types of metadata across various scenarios. Both quantitative and qualitative data were collected. Section 5.1 provides the demographic details of the 56 participants. Section 5.2 outlines the assessment results of using the provenance in dataset search tasks and identifies the provenance information elements needed when searching for datasets. Finally, Section 5.2.3 presents the results of the statistical significance tests.

5.1 Demographics

As outlined in Section 3.2, participants were selected based on their need to search for and reuse datasets, as well as their completion of the survey. The purpose of the study, as described in the same section, was to explore how provenance information can enhance biomedical dataset search and to identify which provenance elements are considered most relevant by researchers. Table 5.1 summarises the participants' demographic and professional backgrounds, highlighting diversity in roles and expertise and reflecting a broad range of experiences within the biomedical research domain.

Domains/Roles	PhD	Postdoc	Research Fellow	Lecturer	Professor	Other	Total
Genomics/Biology	7	2	3	3	6	2	23
Cancer Sciences	3	2	0	0	1	0	6
Biochemistry	2	0	0	2	1	0	5
Neuroscience	1	0	0	0	0	1	2
Immunology	2	1	1	0	0	0	4
Biotechnology	0	2	0	0	1	0	3
Biomedical Engineering	2	0	0	0	0	0	2
Bioinformatics	2	0	0	0	3	1	6
Other	2	0	1	0	0	2	5
Total	21	7	5	5	12	6	56

TABLE 5.1: Demographic information of participants in Study 2

The highest percentage of participants are PhD students, representing 37.5%, followed by professors or assistant professors at 21.4%. Postdoctoral researchers represent 12.5% of the participants, while lecturers and research fellows are both 8.9%. Also, 10% of participants held other roles, including bioinformaticians and MSc students. For instance, one participant from this latter group mentioned that they conducted a research project in clinical neuroscience.

The participants were asked to identify their areas of research. While the specific research domains vary, but all of them fell within the biomedical research field. We identified eight main research domains in the questionnaire: Biology, Genomics, Cancer Sciences, Biochemistry, Biotechnology, Bioinformatics and Neuroscience, with most of our participants being researchers in Biology and Genomics, representing 41.1%. This was followed by Cancer Sciences and Bioinformatics, each one accounting for 10.7%. After that, Biochemistry comprised 8.9%, succeeded by immunology at 7.1%. The domains with the smallest percentage among our participants were Neuroscience, Biotechnology and Biomedical Engineering, which ranged between 5% and 3%.

The respondents were asked to describe their roles when searching for datasets. All participants indicated that they use datasets to conduct biomedical research. Twenty-six participants perform data analysis to investigate medical diagnostics, genetic variants, and more. In addition to conducting research, 11 participants also teach biomedicine students. One participant mentioned teaching biochemistry and biology to clinical laboratory students. Other participants, representing 42%, explicitly emphasised their use of various types of Omics data, such as DNA, RNA, Transcriptomic data and Proteomic data, as well as other types of biological data such as T cells. Based on these results, we consider all participants belong to the target population of this research and we did not exclude any of them.

5.2 Results

This section presents the analysis of the responses to the questionnaire. As mentioned earlier in Section 3.3.2, the questionnaire primarily focuses on presenting different types of metadata across various scenarios. Following each types, the participants were asked to assess the content, and identify the provenance information elements needed when searching for datasets.

5.2.1 Comparison of usefulness of presentation options

As presented earlier in Section 3.3.3, we designed two tasks involving online biomedical dataset searches. For these tasks, participants were provided with four dataset options to select from. The order of the options was altered to investigate whether changing the order would result in any differences. In the first task, the order of options was as follows: Option A, Option B, Option C, and Option D. In the second task, the order of options was as follows: Option C, Option D, Option A, and Option B. The following sections provide explanations of all options.

5.2.1.1 Option A : Dataset metadata

The metadata information in this option includes eight elements, such as Dataset Name, Dataset Repository and Repository ID. Tables 5.2 presents the details of this metadata type. This metadata type was presented as the first option to participants in the first task, while it was the third option in the second task.

For the first task, only 20 participants (35%) chose to download the dataset based on this type of metadata. The participants who only agreed to download this dataset rated the usefulness of each metadata element in making their decision to download the dataset using a Likert-style scale ranging from 4 (Very useful) to 0 (Not useful at all). The mean ratings of the access hyperlink, source organization, data or study type, and dataset name ranged from 3 with a standard deviation of 1.07 to 3.40 with a Standard Deviation (SD) of 0.821, while the other elements had lower mean ratings between 2.20 (1.399 SD) and 2.40 (1.392 SD).

In the second task, 31 participants (55%) agreed to download the dataset. This dataset metadata contains only seven elements, missing the year. Regarding the usefulness judgment of the metadata, the mean ratings of all elements ranged from 3 (1.263 SD) to 2.32 (1.514 SD). Table 5.2 shows a summary of distribution of scores for the metadata elements for both options. All analysis results are presented in Appendix B.

Following this, 65% of participants in the first task and 45% in the second task declined to download the dataset and were unable to make a decision based on the provided metadata. Table 5.3 presents various reasons given by participants, most relate to the lack of detailed information, notably ambiguity regarding the data types within the dataset (P33, P7, P10, P19, P35). For example, P19 stated: *“The type of data is ambiguous — is it expression data, whole genome data?”*.

Another issue raised was the lack of information regarding data size, including sample size. Several participants mentioned the absence of the number of samples in the experiments (P7, P13, P35, P51). Whilst some also stated the importance of knowing the disease types included in the datasets (P10, P13, P51). Additionally, the lack of treatment information, which is essential for decision-making, was highlighted (P13, P46). Further reasons include the absence of accessibility and availability information, and the lack of link to ethics information.

All of the above-mentioned missing information can be considered a lack of provenance information, as they relate to the data history. This finding aligns with (Dixit et al., 2018), which states that the provided information can be incomplete and of low quality, making it difficult to make decisions.

5.2.1.2 Option B : Dataset metadata in abstract

A publication alongside a dataset is commonly presented in biomedicine, providing details about an experiment, including the history of the dataset. Literature repositories, such as PubMed, typically provide a title and an abstract (summary). We selected this option because several researchers use this method to search for datasets, as we discovered in our (previous study 4) and confirmed by (Dong et al., 2017). This type of metadata was presented in our case study as the second option in the first task and the fourth option in the second task.

Regarding the first task, 28 participants (50%) chose to download the dataset based on the provided abstract. The mean ratings measured by the 28 participants for the title and the abstract were close to each other, at 2.89 (1.227 SD) and 2.96 (1.290 SD), respectively. For the second task, the percentage decreased to 42% (24 participants), and the mean ratings in the Likert-style scale of the title and the abstract were 2.43 (1.343 SD) and 3.17 (1.049 SD). Table 5.4 presents the distribution of scores for the title and abstract elements.

Table 5.5 provides several reasons given by our participants for their decision to not download the dataset. As with option A, participants stressed essential information is missing, such as the number of participants or samples used (P2, P7, P18, P45, P54, P51). Another key

Measured Item	Score	Frequency in Task 1 (n=20)	%	Frequency in Task 2 (n=31)	%
Dataset Name	0	1	5%	6	19.35%
	1	0	0%	3	9.68%
	2	6	30%	7	22.58%
	3	3	15%	5	16.13%
	4	10	50%	10	43.30%
Dataset Repository	0	3	15%	4	15.90%
	1	3	15%	0	0%
	2	6	30%	5	16.13%
	3	3	15%	10	32.26%
	4	5	25%	12	38.71%
Repository ID	0	3	15%	3	9.68%
	1	2	10%	1	3.23%
	2	6	30%	2	6.45%
	3	3	15%	10	32.26%
	4	6	30%	15	48.39%
Data Types	0	0	0%	4	12.90%
	1	2	10%	0	0%
	2	5	25%	3	9.68%
	3	4	20%	9	29.03%
	4	9	45%	15	48.39%
Source Organization	0	0	0%	7	22.58%
	1	0	0%	0	0%
	2	7	35%	5	16.13%
	3	4	20%	8	25.81%
	4	9	45%	11	35.48%
Access Conditions	0	3	15%	5	16.13%
	1	1	5%	0	0%
	2	7	35%	4	12.90%
	3	3	15%	4	12.90%
	4	6	30%	18	58.06%
Access Hyperlink	0	0	0%	6	19.35%
	1	0	0%	0	0%
	2	4	20%	3	9.68%
	3	4	20%	5	16.13%
	4	12	60%	17	54.84%
Year	0	Not Provided	Not Provided	5	16.13%
	1			1	3.23%
	2			10	32.26%
	3			8	25.81%
	4			7	22.58%

TABLE 5.2: Helpfulness scores per presentation item for Option A, metadata only

detail missing from the abstract was the description of the methods used in the study (P7, P10, P45). In addition, several participants expressed the need for more specialised information about the data, with gene information being one of the most commonly requested details (P10, P54, P36). Several participants (P7, P54) referred to the lack of ethical information in the abstract. Some participants also missed accessibility details, including criteria to grant access and method for access (P18, P4, P33). A professor (P18) stated: *“Information about this dataset is missing, such as the accession number, type of study, type of data, contact information, and*

Reason	ID	Quotes
Lack of detailed information about the data type	P51	“Doesn’t have enough description about the dataset, e.g. patient cohort, disease types, etc.”
	P10	“It doesn’t show the type of dataset, disease, and tissue etc information.”
	P13	“Not enough description about the dataset, e.g. patient cohort/multi-centre, disease types, stages, treatments, study date etc.”
Lack of treatment information	P46	“I need more specific treatments details.”
	P34	“Get more specific treatments details if they have.”
Lack of accessibility information	P11	“Not enough information about time to download and access conditions, ethics etc.”
	P33	“Doesn’t say if there is phenotype data available or which files are available to download.”
	P7	“What file types are available.”
Data size	P44	“I need more details about data size.”
Ethics	P7	“More information about ethics.”

TABLE 5.3: Quotes of participants that chose not to download the dataset under Option A, metadata only.

cited publications”. Some participants (P7, P54) referred to the lack of ethical information in the abstract. A postgraduate researcher (P45) pointed out the absence of key information in the abstract of the second task, including the year or source.

Measured Item	Score	Frequency in Task 1 (n=28)	%	Frequency in Task 2 (n=24)	%
Title	0	2	7.1%	3	13.04%
	1	1	3.57%	2	8.70%
	2	7	25%	6	26.09%
	3	6	21.43%	6	26.09%
	4	12	42.86%	6	26.09%
Abstract	0	3	10.71%	1	4.17%
	1	0	0%	1	4.17%
	2	5	17.86%	2	8.33%
	3	7	25%	9	37.50%
	4	13	46.43%	11	45.83%

TABLE 5.4: Helpfulness scores for Option B, abstract only

5.2.1.3 Option C : Visual abstract of provenance metadata

This option (visual abstract of provenance metadata) was designed to meet researchers’ needs and assess the effectiveness of provenance information in biomedical dataset searches. As mentioned earlier in Section 3.3.1, this graph, shown in Section 3.6, was designed to present provenance information, from obtaining ethical approval to storing the outcomes. Provenance information was extracted in the form of PROV-DM components and presented in a visual abstract. This

Reason	ID	Quotes
Lack of detailed information about the experiment	P10	"The abstract didn't mention any scientific background like what sample, patient, type of data, genes information."
	P45	"Epidemiologic information, number in each age group, sex, onset of disease, etc."
	P51	"The abstract didn't mention what sample size, patient, type of dataset, etc."
	P2	"To expand the subject of the study, i.e., samples used."
Lack of accessibility information	P4	"Is this publicly available, how strict are the access criteria, is it a simple email to get the data or will I need to go through contracts and RIS. For long term projects closed access data is ok, but for short-term I'd avoid any closed access data as it can take months upon months"
	P33	"This doesn't have any information regarding the data, where it is available, just seems like a review paper."
	P7	"What file types are available."
Lack of ethics information	P54	"Nothing about ethics."
	P7	"I need more information about ethics."

TABLE 5.5: Quotes of participants that chose not to download the dataset under Option B, abstract only.

type was shown as the third option to participants in the first task, while it was the first option in the second task.

In the first task, 46 participants, representing 82.14% of the total, agreed to download the dataset, indicating positive significant difference compared to the previous types of dataset metadata. Next, we measured the scores provided by all the participants who selected this approach, for evaluating the usefulness of the provided provenance information, the relationships between the steps, and other general information, such as authors and year. The mean ratings results were similar to the above metadata types, but in fact, they varied depending on the number of participants who chose each type. The mean ratings for both elements were 3.17 and 3.48 (1.338 SD and 0.888 SD), while for the other elements were less than 2.90 (1.269 SD).

Regarding the second task, 45 (80.36%) participants chose to download this dataset. For the usefulness judgment, the mean ratings of all elements ranged from 2.67 (1.552 SD) to 3.22 (1.106 SD). Table 5.6 presents the scores for the above information. Initially, we noticed that this option surpassed the two options mentioned above. As with the previous options, we asked participants who did not download the datasets based on this option for suggestions to improve it prior to the extractor's implementation.

A small number of participants, around 10% in both tasks, selected not to download the dataset based on this information provided, and 8% were unsure. The primary reason was the desire to explore more dataset options. A participant (P11) mentioned: *"you can't know whether you have a good data set until you've reviewed all those available and seen which ones use"*.

Another reason was associated with accessibility information as mentioned by (P4, P33). Table 5.7 provides a summary of the reasons given by our participants.

Measured Item	Score	Frequency in Task 1 (n=46)	%	Frequency in Task 2 (n=45)	%
Title + Authors + Year	0	4	8.7%	7	15.56%
	1	5	10.87%	5	11.11%
	2	9	19.57%	6	13.33%
	3	7	15.22%	5	11.11%
	4	21	45.65%	22	48.89%
Provenance Information Details	0	1	2.17%	3	6.67%
	1	0	0%	0	0%
	2	6	13.04%	5	11.11%
	3	8	17.39%	13	28.89%
	4	31	67.37%	24	53.33%
Relationships Between the Steps	0	5	10.87%	3	6.67%
	1	1	2.17%	3	6.67%
	2	4	8.70%	5	11.11%
	3	7	15.22%	11	24.44%
	4	29	63.04%	23	51.11%
Explanatory Icons	0	4	8.70%	4	8.89%
	1	2	4.35%	3	6.67%
	2	9	19.57%	8	17.78%
	3	11	23.91%	8	17.78%
	4	20	43.48%	22	48.89%

TABLE 5.6: Helpfulness scores for Option C, provenance information.

Reason	ID	Quotes
Need to see more dataset options	P11	"You need an idea of all options."
	P37	"To get more options."
	P41	"You can't know whether you have a good data set until you've reviewed all those available and seen which ones others use."
Lack of accessibility information	P4	"Closed or open access."
	P33	"No indication of which data is available to download."

TABLE 5.7: Quotes of participants that chose not to download the dataset under Option C, provenance information.

5.2.1.4 Option D : Dataset provenance metadata combined with abstract

This option was designed to present both the abstract of a published paper (option B) and provenance information (option C). The graph in this option included the provenance information and incorporated the abstract of the paper published based on the datasets. It aimed to explore whether this combination would enhance the dataset search compared to the previous option.

A copy of the survey is provided in Appendix B, where this option was presented as the fourth option to participants in the first task and the second option in the second task.

When combining options B and C, 87% of our participants, followed by 78% of our participants chose to download the dataset in both tasks. In evaluating the usefulness of the provided information, the mean ratings for all elements in the first task were above 3 (all SD were less than 1.296 SD), except for the abstract, which rated at 2.89 (1.323 SD).

In the second task, the mean ratings for all elements were around 3 (1.149 SD), except for the abstract, which rated at 2.67 (1.426 SD). Table 5.8 presents the scores for the above information. However, the only reason mentioned for not downloading this dataset was the lack of information about ethics (P7, P54), as shown in Table 5.9.

Measured Item	Score	Frequency in Task 1 (n=49)	%	Frequency in Task 2 (n=44)	%
Title + Authors + Year	0	4	8.51%	6	13.95%
	1	3	6.38%	2	4.65%
	2	10	21.28%	10	23.26%
	3	7	14.89%	7	16.28%
	4	23	48.94%	18	41.86%
Provenance Information Details	0	1	2.08%	2	4.55%
	1	0	0.0%	1	2.27%
	2	6	12.50%	5	11.36%
	3	10	20.83%	12	27.27%
	4	31	64.58%	24	54.55%
Relationships Between the Steps	0	2	4.08%	2	4.44%
	1	3	6.12%	3	6.82%
	2	3	6.12%	9	20.45%
	3	13	26.53%	12	27.27%
	4	28	57.14%	18	40.91%
Explanatory Icons	0	5	10.42%	5	11.36%
	1	2	4.17%	1	2.27%
	2	3	6.25%	5	11.36%
	3	15	31.25%	14	31.82%
	4	23	47.92%	19	43.18%

TABLE 5.8: Helpfulness scores for Option D, provenance and abstract.

Reason	ID	Quotes
Lack of accessibility information	P54	"Again, nothing about ethics."
	P7	"I want to see more information about ethics."

TABLE 5.9: Quotes of participants that chose not to download the dataset under Option D, provenance and abstract.

5.2.2 Information gained from provenance

As explained in Section 3.3.3, a multiple-choice question was included under options C and D to identify the specific pieces of relevant information gained by participants when examining provenance. We can then estimate whether the presented data provenance is sufficient to help researchers select datasets or if it requires improvement. Based on this, we can identify the necessary provenance elements for the extractor's implementation. Under this question, six options were designed based on several existing works from the literature (Johns et al., 2023), as well as semi-structured interviews conducted with biomedical researchers in Section 4.2.8 as part of this PhD study to determine the requirements for dataset search: a) obtaining ethical approval; b) how the data was collected; c) how the data was processed; d) understanding the steps followed; e) understanding the outcome of the experiment; f) Other, allowing a participant to provide input.

On average, 67.03% of the participants, who intended to download the datasets after viewing an option including provenance, were able to obtain information about the ethical approval of the data included in the datasets, while 32.97% could not find enough information about this part. This result is the lowest percentage among all elements of the provenance information. This could be because published papers often lack comprehensive information about ethics we observed, most papers only provide the ethical approval number.

Over 84% of the participants could understand how the data was collected and processed from the proposed provenance information. To understand the outcome of the experiment that the dataset is based on, more than 88% of the participants declared to have understood this element from the provided provenance information. Finally, regarding the entire procedures followed in the experiment, only 2.17% of participants could not obtain it from the provenance information. Overall, the results indicate participants were generally satisfied with the presented provenance information.

For the second task, the followed procedure had the highest percentage (100%). The lowest percentage of obtaining information about ethical approval was (66.67%). In this task, we observed that it outperformed the first task. Table 5.10 presents a summary of the frequency and percentage of all elements in both tasks.

5.2.3 Statistical Significance Tests

In this section, we present the data analysis derived from the questionnaire. This section includes four types of statistical tests, each examining and analysing the relationships between different

Information Gained	Response	Frequency in Task 1	%	Frequency in Task 2	%
A) Obtaining ethical approval	Yes	31	67.39%	30	66.67%
	No	15	32.61%	15	33.33%
B) How the data was collected	Yes	39	84.78%	38	84.44%
	No	7	15.22%	7	15.56%
C) How the data was processed	Yes	39	84.78%	41	91.11%
	No	7	15.22%	4	8.89%
D) Understanding the steps followed	Yes	45	97.83%	45	100%
	No	1	2.17%	0	0%
E) Understanding the outcome of the experiment	Yes	41	89.13%	40	88.89%
	No	5	10.87	5	11.11%

TABLE 5.10: Summary of provenance information gained

aspects of the collected data. The analysis employed Chi-square tests, Fisher's Exact tests, t-tests, ANOVA tests, and Cronbach's alpha test.

5.2.3.1 The differences between showing existing metadata and showing provenance information

One of the most important aspects of this study is to determine the effectiveness of using provenance information in dataset search. To compare the differences in the means between the existing dataset metadata types and the proposed ones, paired or dependent samples t-test was used (Okoye and Hosseini, 2024).

Firstly, we compared the dataset metadata (option A in Section 5.2.1.1) with the provenance information (option C in Section 5.2.1.3). The test results showed a negative t-value (-0.425) and a negative mean difference between the two options (-0.482), indicating that the mean of option A is lower than that of option C. This t-test result demonstrates a statistically significant difference between the means of these two approaches at a 95% confidence level ($p = 0.001$, <0.05). Therefore, the negative value confirms a statistically significant difference between these options, with the metadata (option A) being significantly less effective than the dataset description using provenance information (option C).

Secondly, we compared the option of the abstract (option B in Section 5.2.1.2) to the dataset description combining provenance information and the abstract (option D in Section 5.2.1.4). The result shows a negative t-value (-4.078) and a negative mean difference between the two options (-0.446), indicating that the mean of (option B) is lower than that of (option D). The p-value confirms that there is a statistically significant difference with a 95% confidence level ($p = 0.001$, <0.05).

To ensure the accuracy of our analysis, we conducted another test. A Univariate ANOVA, a statistical test used to compare the means of at least three different groups (Kim, 2014), was applied to determine if there are statistically significant differences between the groups.

In our study, we compared the means of all options provided to identify any significant differences between them. All the four options were included in this test. The results indicated that there are statistically significant differences among the options, meaning that at least one of the four options is significantly different from the other three options.

The test results indicate that the means are significantly different from each other, in which the p-value is <0.001 and the F-statistic = 13.234. The F-statistic indicates that the variation between the options is 13 times larger than within the other options, and the p-value is ($p < .001$), which provides strong evidence to confirm this.

5.2.3.2 Association between roles and dataset metadata

One of our interests is to determine whether there is a relationship between the participants roles and their selection of dataset metadata. The aim is to discover if selecting a dataset based on provenance information requires a certain level of experience. For instance, are postgraduate researchers able to read and understand the provenance information, or does it require a higher level of experience?

We identified a common test for this analysis: the Chi-Square test, which are typically used to find relationships between two categorical variables, in this case, roles and responses in each approach (Franke et al., 2012). The equation of this test: was performed as

$$\chi^2 = \frac{1}{d} \sum_{k=1}^n \frac{(O_k - E_k)^2}{E_k}$$

where O is the observed values, and E is expected values. We should note that the p-value indicates the probability that the observed result occurred by chance alone. For example, a p-value of 0.05 indicates a 5% probability that the result occurred by chance. When $P < 0.05$, it means there is a 95% likelihood that the result did not occur by chance. Therefore, there will be a significant relationship statistically (Field, 2024).

In our study, these tests were applied to determine whether there is a relationship between participants' roles and the types of dataset metadata selected. All responses in both tasks were classified according to the participants' roles. For these results to be considered statistically significant, the p-value must be less than 0.05. As shown in Table 5.11, all p-values for the

relationship between roles and dataset metadata approaches were greater than 0.05, indicating no association between these variables. This suggests that selecting a dataset based on provenance information can be effectively done by biomedical researchers in different roles. In addition, it indicates that provenance information can be understood by biomedical researchers with varying levels of experience.

Metadata Approaches	Task (T)	Test	p-value
Dataset description in metadata	T1	Chi-square	0.631
	T2	Chi-square	0.571
Dataset description in abstract	T1	Chi-square	0.473
	T2	Chi-square	0.415
Dataset description in provenance information	T1	Chi-square	0.213
	T2	Chi-square	0.320
Dataset description in provenance information with abstract	T1	Chi-square	0.360
	T2	Chi-square	0.370

TABLE 5.11: Pearson Chi-square test results for the association between roles and different dataset description approaches across tasks.

However, we found that some cells are less than five values, indicating a potential issue with the Chi-Square test. The Chi-Square test is suitable for large sample sizes; therefore, to obtain more reliable results, we conducted Fisher's Exact Test, which is more appropriate for small sample sizes (Bower, 2003). All p-values from Fisher's Exact Test confirm that there is no association between the roles of the participants and their selection, as shown in 5.12.

Metadata Approaches	Task (T)	Test	p-value
Dataset description in metadata	T1	Fisher's Exact	0.753
	T2	Fisher's Exact	0.716
Dataset description in abstract	T1	Fisher's Exact	0.533
	T2	Fisher's Exact	0.887
Dataset description in provenance information	T1	Fisher's Exact	0.440
	T2	Fisher's Exact	0.372
Dataset description in provenance information with abstract	T1	Fisher's Exact	0.168
	T2	Fisher's Exact	0.619

TABLE 5.12: Fisher's Exact test results for the association between roles and different dataset description approaches across tasks

5.2.3.3 Association between metadata options and responses

Another analysis was conducted to determine if there is an association between the metadata options and the participants' responses, or if the responses were random. We observed in our

qualitative analysis that the selection of options C and D excelled for options A and B; thus, we had to test this statistically. We conducted this analysis to confirm that the selection of metadata options was not random and corresponded to our qualitative analysis.

First, we applied the Chi-Square Test, as explained in the previous section. The p-values for both tasks were $p < 0.001$, indicating that there is a statistically significant association between these two variables. With this test, we obtained a notice that there were “4 cells (33.3%), which have an expected count less than 5. The minimum expected count is 4.00”. This suggests that the Chi-Square test may be less reliable. In addition, this test is generally more reliable with larger sample sizes. Therefore, to obtain more accurate results and ensure the quality of our analysis, we searched for other possible techniques.

To further verify the results, we also conducted the Fisher’s Exact Test. The p-values from both tasks were also $p < 0.001$, confirming the statistically significant association. Table 5.13 presents the result of both tests. Screenshots of the SPSS analyses and tests are presented in Appendix B.

Tests	Task (T)	Value	Significance (2-sided)	Exact Significance (1-sided)
Pearson’s chi-squared test	T1	48.322	<.001	<.001
	T2	24.606	<.001	<.001
Fisher’s exact test	T1	49.230		<.001
	T2	24.196		<.001

TABLE 5.13: Statistical test results for the approaches across two tasks

5.2.3.4 Reliability tests

We conducted a reliability analysis to assess whether the variables in each type of metadata were measured adequately by the scale. We asked the participants to rate the usefulness of each element in helping them decide whether to download the dataset. Figure 3.7 presents a sample of the Likert-style scale used. The scale is defined in fig 5.1:

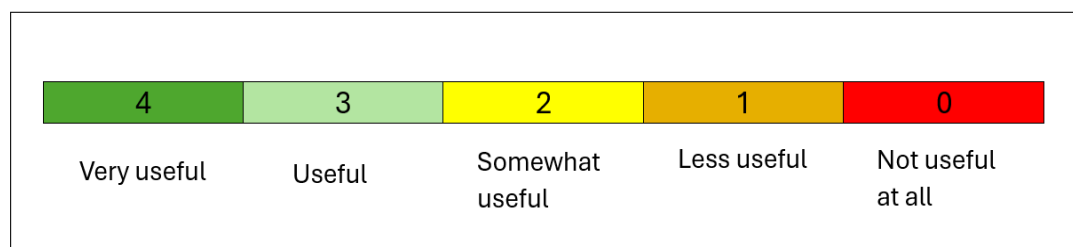


FIGURE 5.1: Likert-style scale for the usefulness assessment.

For this study, we analysed the questions that used a five-point Likert scale, consisting of a total of 35 items across different questions, by applying Cronbach’s alpha. It aims to measure the

consistency of a scale or test (Tavakol and Dennick, 2011), which is expressed as a value between 0 and 1. A value of 0.7 in this test is considered sufficient to indicate high internal consistency. In both tasks, all values ranged between 0.760 and 0.892 across all items, indicating acceptable to good internal consistency, as shown in Table 5.14.

Metadata Approaches	Task (T)	No. of Items	Cronbach's Alpha
Dataset description in metadata	T1	7	0.773
	T2	8	0.892
Dataset description in abstract	T1	2	0.838
	T2	2	0.760
Dataset description in provenance information	T1	4	0.778
	T2	4	0.817
Dataset description in provenance information with abstract	T1	4	0.840
	T2	4	0.818

TABLE 5.14: Cronbach's alpha analysis

5.3 Discussion

Our results demonstrate that using provenance metadata in dataset search significantly enhances users' ability to assess biomedical datasets. Additionally, provenance metadata outperformed the currently used types of metadata in Datamed and PubMed, as it provides more detailed information about datasets, which enables users to better assess them. In both tasks, over 82% of participants preferred provenance information, with this percentage increasing to 87% when provenance was combined with the abstract. In addition to this, the statistical significance tests demonstrate a significant difference between the existing approaches and the proposed ones, which include provenance information. This explains the significance of the provenance information when searching for dataset search, particularly biomedical datasets. By comparison, none of the other options exceeded 55% in either task. A key reason for not selecting one of these two options (A and B) is the lack of detailed information. These findings align with prior research (Dixit et al., 2018), which highlights the challenges associated with incomplete or low-quality metadata when assessing datasets online.

By measuring the improvements achieved through the use of provenance metadata in biomedical dataset search (ranging from 67% to 100%) across several key aspects, including ethical approval information, data collection procedures and study outcomes, this study provides a detailed empirical assessment of its impact. Furthermore, the statistical tests, such as the Chi-Square test and the Fisher's Exact test and sample t-tests, revealed statistically significant differences between

existing metadata approaches in two biomedical portals and provenance metadata approaches, suggesting that the proposed provenance information extracted from publications associated with dataset can enhance biomedical dataset search.

The suitability of understanding provenance information across different levels of expertise was an important aspect explored in this study. The aim was to meet the FAIR principles by integrating provenance information into biomedical dataset search. Additionally, there was a concern that understanding provenance information might require a high level of expertise; therefore, we investigated this by conducting a statistical test. The Chi-Square test indicates that there is no relationship between the participants' roles and their selection of the provenance metadata options (option C and option D), suggesting that the proposed provenance information can be understood by biomedical researchers of all levels.

Our results suggest that providing provenance information in a visual representation during dataset search reduces uncertainty in dataset assessment, and enhances online dataset search for biomedical research domains. Here, provenance information can assist biomedical researchers in gaining a deeper understanding of the history of the data within datasets, aiding in assessment and decision-making during dataset search. This improvement in dataset search could contribute to reducing researchers' time, enhancing the process of understanding and evaluating datasets, and mitigating resource usage by avoiding the need to conduct research from scratch. These findings align with [Sahoo et al. \(2023\)](#)'s findings on the use of the FAIR principles in dataset search, demonstrating how provenance information can enhance dataset evaluation and usability through the application of these principles.

However, certain challenges remain. A drawback of the activity-centered provenance graph is that it cannot be automatically generated. Consequently, we created a graph in this study, as we could not set up an experiment where subjects could independently search a repository containing both relevant and irrelevant datasets. Moreover, it is challenging to convince paper authors to complete structured metadata, and it may be difficult to encourage them to prepare a provenance graph (unless forced by venues/publishers).

5.4 Summary

The aim of this study is to assess the effect of providing provenance metadata when searching for biomedical datasets. Our findings demonstrate the considerable effectiveness of presenting provenance metadata in a visual diagram, with over 45 participants selecting datasets based on the provenance information. Provenance information enabled most of our participants to gain

information about data collection, data processing, the steps followed and the final outcome. The piece of information that was least available was ethical approval related to the data within the dataset, as reported by our participants. This study contributes to the information retrieval domain by providing empirical evidence of the effectiveness of providing provenance metadata in dataset search. The integration of provenance information into dataset search can potentially improve the dataset search domain across all scientific disciplines. Future user studies could compare the effectiveness of different provenance visualisations vis-a-vis the complexity of their generation or if an interactive interface provides additional benefits over flat representations.

Chapter 6

Provenance Information Extractor

In this chapter, we present the extractor, a tool developed to extract provenance information from biomedical articles. This tool was built using GPT-4o (Hurst et al., 2024), a generative pre-trained transformer, via its API. The extractor was specifically designed to address RQ4. Section 6.1 describes the extractor's implementation, architecture, and components. Section 6.2 presents an experiment aimed at developing and assessing prompt patterns for provenance information extraction. In Section 6.3, we explain and discuss a scalability experiment conducted for the extractor.

6.1 Implementation of Extractor

In this section, we present the implementation of our extractor. The extractor builds upon an existing LLM, which is invoked using its API. It uses various tools and data formatting techniques to process data, starting with its retrieval from storage and ending with the preservation of provenance information back into the storage system. Figure 6.1 presents a high-level overview of our extractor architecture. It consists of three main levels and several key components, which are subsequently explained.

The provenance extractor was implemented in Python, because of its comprehensive standard libraries. Figure 6.2 illustrates the workflow of the extractor, describing all the steps involved in its implementation. Several libraries were utilised in this tool, including:

- **requests**¹: The first used library requests, which is utilised to make HTTP requests. This library is appropriate for fetching and sending requests to APIs.

¹https://www.w3schools.com/python/module_requests.asp

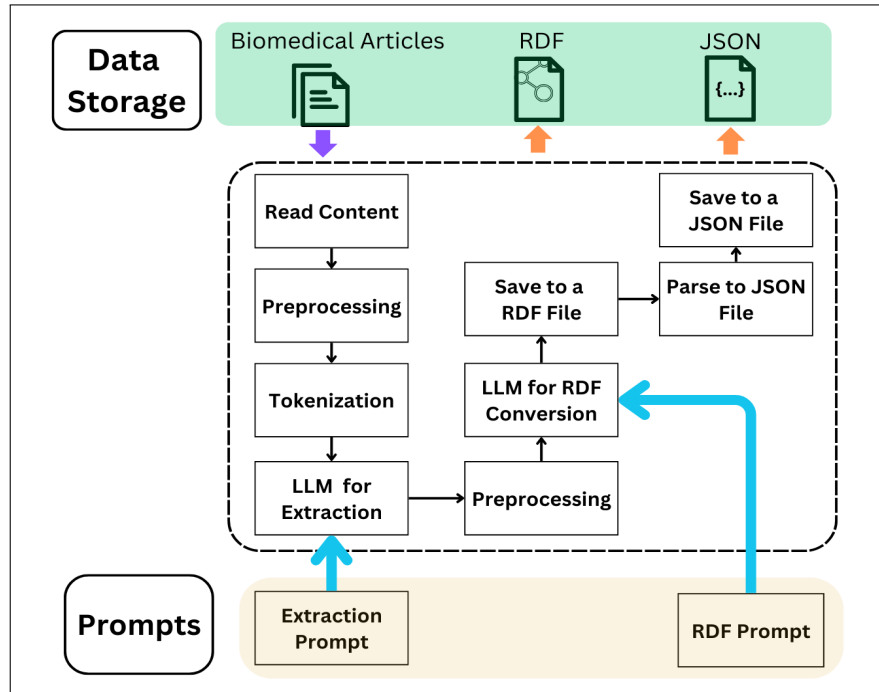


FIGURE 6.1: Architecture of the extractor.

- `json`²: It is an abbreviation for JavaScript Object Notation, a data exchange and storage format inspired by JavaScript. JSON in Python is a built-in package that can be utilised to deal with JSON data.
- `fitz` (PyMuPDF)³: This module connects MuPDF to Python. It is used to read, extract, and deal with data from PDF files while preserving the main PDF layout (Koning, 2022).
- `os`⁴: It uses to interact with files' operating system, such as creating, deleting files or folders.
- `re`⁵: It supports regular expressions for pattern matching operations, enabling developers to define string matches using regular expression.
- `nltk`⁶: It supports machines in handling natural language tasks. This library includes packages for text analytics, such as tokenization and lemmatization.
- `tiktoken`⁷: This is an effective tool for estimating the number of tokens and converting text into tokens. This library is crucial for working with LLMs, such as GPT. It enables developers to understand the tokenisation process and estimate the cost of text input.

²https://www.w3schools.com/python/python_json.asp

³<https://pymupdf.readthedocs.io/en/latest/tutorial.html>

⁴https://www.w3schools.com/python/module_os.asp

⁵<https://docs.python.org/3/library/re.html>

⁶<https://www.nltk.org/>

⁷<https://github.com/openai/tiktoken>

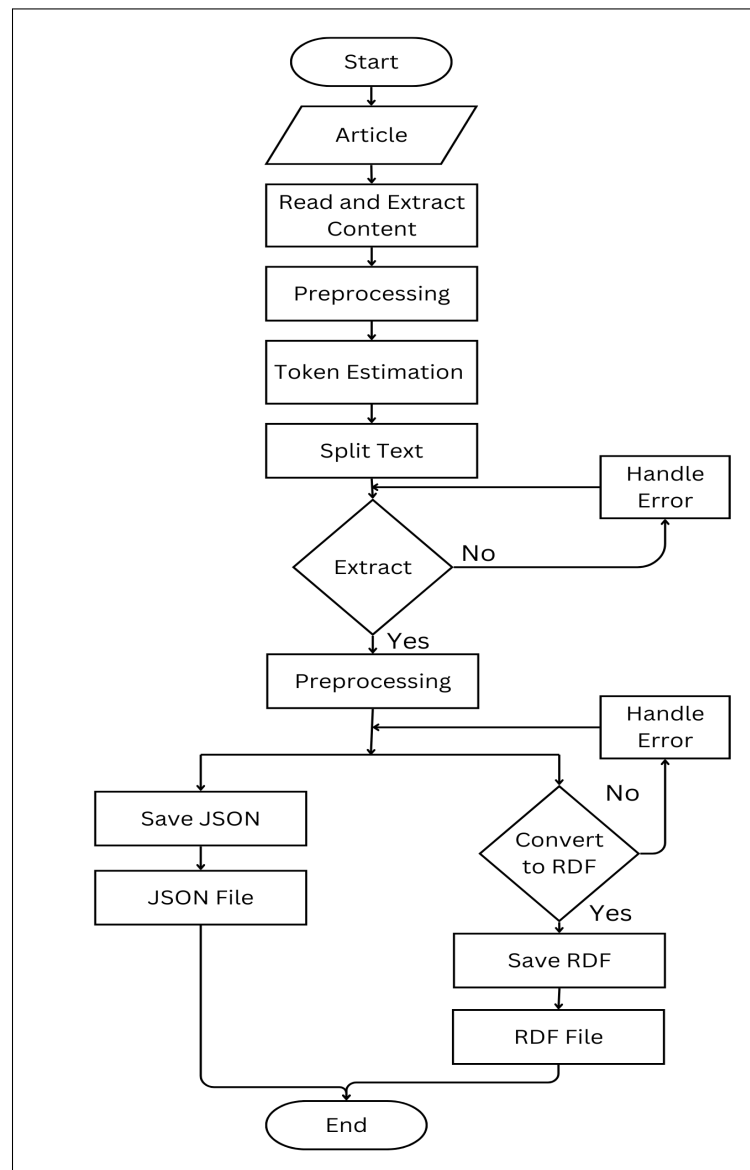


FIGURE 6.2: The extractor workflow diagram.

- `time`⁸: This provides developers with several temporal functions, such as completion time and timestamps.

To start sending requests and receiving responses to the OpenAI API, which is explained in Section 6.1.2, we used a headers dictionary to communicate with it, which can be invoked using the `requests` library. This dictionary is typically used to provide metadata about the user's requests, such as the content type.

Thereafter, we used the `fitz` library to extract and read text within PDF files, employing several functions, such as `[fitz.open]`, `[page.get_text]`, and `[doc.load_page]`. Due to the length

⁸<https://docs.python.org/3/library/time.html>

of PDF files and rate limit issues, we used the `[tiktoken]` library to estimate the token count required for the texts within each PDF file. If the token count exceeded the rate limit, the texts would be split into paragraphs and sentences using several functions in the `[nltk.tokenize]` library, such as `[sent_tokenize]`.

We cleaned the content as it is a crucial phase in data analysis and provides accurate and consistent data without errors. This step was conducted to prepare the content to be sent for provenance extraction using GPT-4o as well as to prepare the extracted provenance information for storing, since we observed extra symbols. For this step, we used the `re` library which includes several functions, such as `[re.sub]` that were used to remove extra `*` and `+`.

Subsequently, the whole text for each paper and the provenance extraction prompt, as explained in Section 3.4.2 were sent to the model. Several mechanisms were implemented to address encountered errors and prevent their recurrence, including verifying successful HTTP requests and addressing rate limit errors. Figure 6.3, for example, illustrates a rate limit error encountered during code testing, which was resolved using the `[get]` method provided by Python dictionaries to handle such errors.

The extracted provenance information in PROV-DM was stored in JSON format using the `json` library. We used the JSON format due to its flexibility in integration with database systems and query languages (Ong et al., 2014). Effective use of data stored in any format, including JSON, requires correct data modelling. This is a fundamental component in data management field involving various tasks, such as data searching, indexing and integration (Pezoa et al., 2016). Additionally, we asked the model to convert the results to RDF to be used for evaluation purposes. RDF, an abbreviation for Resource Description Framework, is a W3C standard developed to model web objects as part of the Semantic Web (Özsu, 2016). For instance, DBpedia can extract information from Wikipedia and store it in RDF format (Bizer et al., 2009). RDF is characterised by several key aspects, including variety, veracity, velocity and volume (Özsu, 2016). In this study, we adopted this format to store the extracted provenance information in RDF for further analysis.

6.1.1 Output Validation

Data validation is an integral part of the data process, which is essential for exploring and enhancing data quality (Xie et al., 2017). The importance of data and format validation is closely connected to other data management tasks, such as data collection and processing (Martin et al., 2008). Consequently, we validated the extracted provenance information and the formats used

```
Failed with status code 429 for Genetic.pdf
Response: {
  "error": {
    "message": "Rate limit reached for gpt-4o in
organization org-FWlrOUumYf27qq63pnC6VbWS on tokens per min (TPM): Limit 30000,
Used 19624, Requested 12421.
Please try again in 4.09s. Visit https://platform.openai.com/account/rate-limits to learn more.",
    "type": "tokens",
    "param": null,
    "code": "rate_limit_exceeded"
  }
}
```

FIGURE 6.3: Example of a limit error message.

to ensure consistency and quality. Ten output results were randomly selected to perform this validation. This output was a provenance information in PROV-DM data, which stored in RDF and JSON documents.

To check the validity of the RDF output, we used the W3C RDF Validation Service⁹, which is designed to validate and examine various aspects of RDF data, including syntax and structure. It can also convert and visualise the data into a graph of triples. Figure 6.4 presents a sample of RDF validation using the above service.

To validate the provenance data included in the files, we utilised ProvStore, an online repository that supports provenance documents. This repository provides various services, including browsing, visualising, validating, storing and managing provenance data (Huynh and Moreau, 2014). ProvValidator¹⁰, a service provided within ProvStore, is used to verify the validity of PROV representations or translate them into alternative representations. Figure 6.5 presents a sample of provenance validation using the above service.

6.1.2 ChatGPT API

OpenAI, like many other services, provides users with an Application Programming Interface (API). The purpose of this is to facilitate the invocation of its services and enable integration into existing models, such as GPT-4o. Through this integration, developers can obtain natural

⁹<https://www.w3.org/RDF/Validator/>

¹⁰<https://openprovenance.org/service/validator.html>

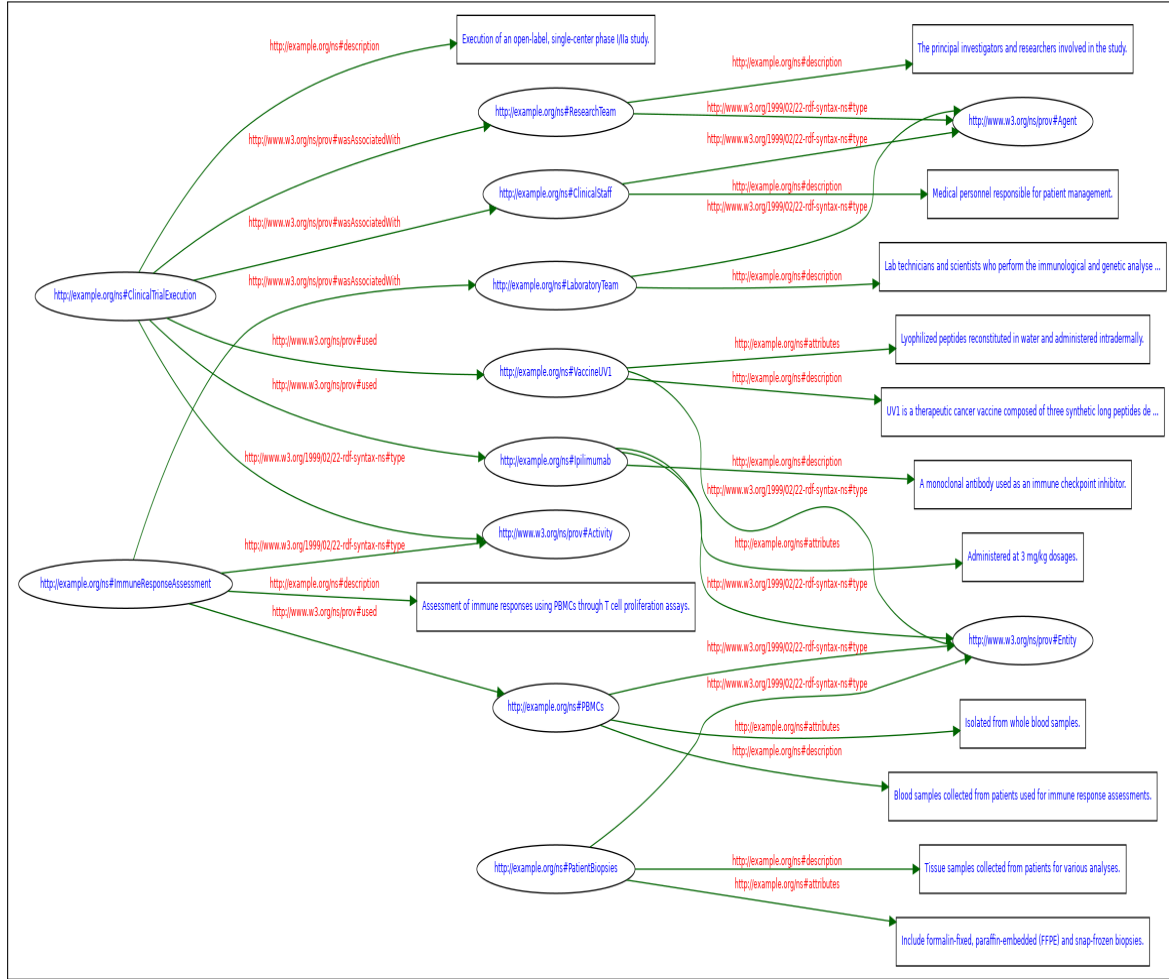


FIGURE 6.4: Example of RDF file validation.

language responses from the models by sending queries. To benefit from these services, users or companies must understand the API's pricing model, which basically depends on the number of tokens used in each request.

Tokenisation is an essential procedure in the process of LLMs (Islam and Moushi, 2024; Lappalainen and Narayanan, 2023). The purpose of this step is to divide input texts and output results into smaller units, known as tokens (Briganti, 2024). Tokens consist of a set of segments, such as letters, words, or symbols, based on the tokenization scheme method used in the LLM (Lappalainen and Narayanan, 2023). The quality of the results produced by the model can be influenced by tokenization process.

GPT-4o introduces several enhancements across various aspects, such as improved tokenization efficiency and multimodal capabilities (Islam and Moushi, 2024). Additionally, it is characterised by a next-generation transformer architecture, which enables a deeper understanding and the generation of relevant responses. The rate limit of this model is five times higher more than tha

No guarantees are provided as to the correctness of this validation result.

☒ Ordering Constraints (0)

☒ Merge Constraints (0)

☒ Type Constraints (0)

☒ Malformed Expressions (0)

☒ Specialization Constraints (0)

☒ Duplicate Statements (0)

☒ Misleading Qualified Names (0)

PROV-Constraints (<http://www.w3.org/TR/prov-constraints/>)

communication-generation-use-inference (<http://www.w3.org/TR/prov-constraints/#communication-generation-use-inference>)

FIGURE 6.5: Provenance information validation.

of GPT-4 Turbo in terms of tokens per minutes, due to its new o200k base tokenizer algorithm. By using this algorithm, the model improves semantic coherence in produced texts and enhances the processing of multiple languages (Islam and Moushi, 2024).

6.2 Provenance Extraction Prompts Experiment

The objective is to compare the accuracy of each prompt pattern by employing a systematic approach, as discussed in Section 3.4.3.2. This approach was applied to interact with LLMs using all the different prompt patterns outlined in the previous section, to identify the most appropriate one for our purpose.

Evaluating Prompt Steps 1 and 2 utilised a scoring scale for the intermediate prompts: 2 if the prompt answered correctly on the first time, 1 if the prompt answered correctly in multiple attempts (up to 5 attempts) and 0 if no prompt answered correctly in multiple attempts (up

to 5 attempts). ChatGPT-4o was tasked 48 times with accessing, extracting and analysing the six biomedical papers, demonstrating its effectiveness in information extraction from academic papers. The accuracy percentage ranged from 96.67% to 68.89%, as shown in Table 6.1, where at least three out of five prompts from the prompt set being answered correctly in both the first and second steps of the interaction.

Paper	Paper Length	Format Complexity	Avg. Accuracy
1	11	Standard	96.67%
2	8	Low	81.11%
3	14	Standard	82.22%
4	8	Standard	82.22%
5	11	Standard	75.56%
6	9	High	68.89%

TABLE 6.1: Average Percentage of Accuracy in Responses.

Evaluating Prompt Steps 3 and 4 compared the results of LLM-based extraction against expert-created provenance in PROV-DM components (entities, activities, agents, and relationships) for each paper. We measured the quality of the LLM's output using standard information retrieval metrics and compared it to our expert-generated provenance information, as described in Section 3.4.3.3. The results are presented in Table 6.2. For example, suppose the provenance returned by the LLM (P) contains the set of these entities $\{e, a, c, r\}$, extracted from the uploaded paper. The expert-generated provenance (Pe) contains the set $\{e, b, c, i, r\}$. When comparing (P) to (Pe), we consider the matching elements $\{e, c, r\}$, which results in a precision of $3/5 = 0.6$.

Prompt patterns G and C exhibited the highest precision, at approximately 0.901 and 0.890 respectively, for the average of all components (Entities, Activities, Agents, Relationships). In contrast, the lowest average precision appeared in pattern B, with a score of around 0.798. Turning to the average recall for all components, as shown in Table 6.3, patterns D and H demonstrated the highest average recall, with scores of approximately 0.66. However, the lowest average recall is seen in Pattern E, with a score of around 0.533.

6.3 Scalability Experiment

The experiment was conducted between 18 June and 8 July 2024. We used a Dell Latitude 5410 running Windows 10 Enterprise (Version-22H2). The configuration of the device was as follows: Intel(R) Core(TM) i5-10310U CPU @ 1.70GHz, 2.21 GHz processor, 16 GB RAM, 64-bit operating system, and a PC SN530 NVMe WDC 256GB. Anaconda Navigator 2.6.1 and Notebook 7.0.8 were used to write the Python script.

	Avg for Entities	Avg for Activities	Avg for Agents	Avg for Relations	Avg for all Provenance
A	0.944	0.906	0.598	0.763	0.803
B	0.905	0.905	0.623	0.761	0.798
C	0.847	1	0.835	0.876	0.889
D	0.88	0.94	0.708	0.743	0.817
E	0.971	1	0.751	0.746	0.8675
F	0.946	0.965	0.725	0.778	0.853
G	0.975	0.958	0.8483	0.823	0.901
H	0.958	0.9583	0.805	0.7667	0.872

TABLE 6.2: Average Precision For All Prompt Patterns.

	Avg for Entities	Avg for Activities	Avg for Agents	Avg for Relations	Avg for all Provenance
A	0.450	0.565	0.635	0.65	0.575
B	0.378	0.678	0.625	0.6	0.570
C	0.475	0.648	0.725	0.601	0.612
D	0.455	0.795	0.695	0.706	0.66
E	0.398	0.581	0.585	0.566	0.532
F	0.446	0.693	0.571	0.71	0.605
G	0.465	0.696	0.668	0.513	0.585
H	0.49	0.7433	0.726	0.681	0.66

TABLE 6.3: Average Recall For All Prompt Patterns.

To monitor the extractor’s performance, we used the `time` library to record the response time for each process, defined as the time taken to process each file per second. We also recorded the total time spent, from uploading the files to obtaining the extracted provenance information at the end of each round, per second. Additionally, we tracked any errors or failures that occurred as the file volumes increased. The aim of this step was to estimate the time required to extract all exome sequencing experiments in PubMed.

To measure the cost of using the API and estimate the cost for this PubMed files, we monitored the usage cost provided by OpenAI for each iteration performed. Furthermore, we needed to understand various rate limits, including numbers of tokens allowed per minute, numbers of requests per minute or day and usage tiers, which may impact the rate limits and affect the performance of the experiment. We measured the cost at the end of each iteration to perform estimations. OpenAI provides the cost in dollars (\$); therefore, all cost estimations were reported in \$.

To ensure the scalability of our extractor, we conducted several iterations with varying dataset sizes and measured the response time and cost for each scenario. In this process, the dataset size (number of files) was gradually increased, allowing us to observe the extractor’s performance in

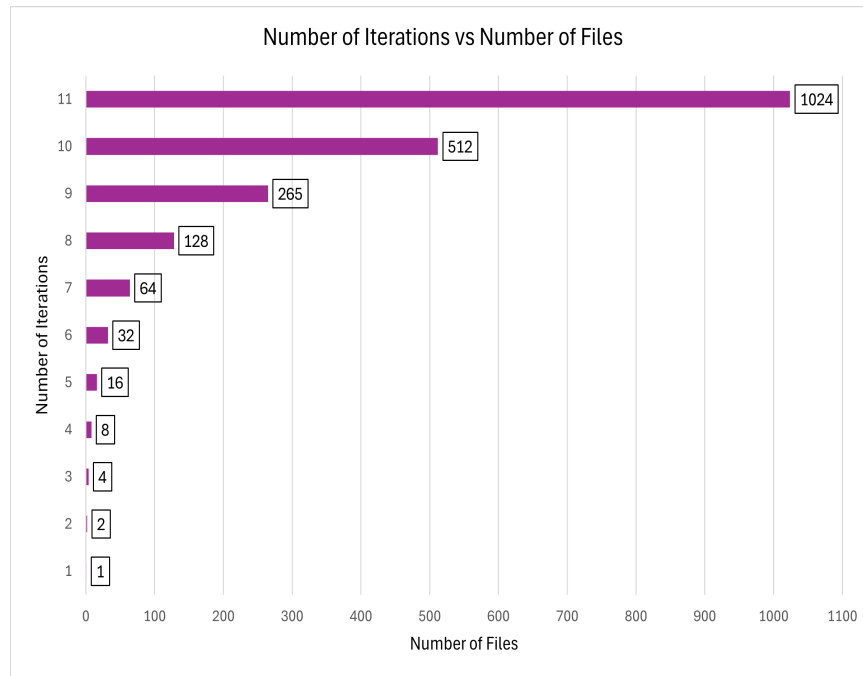


FIGURE 6.6: Number of iterations and number of files

terms of time and cost. We started with 1 file and then doubled the number of files 11 times, ultimately reaching a total of 1,024 files. Figure 6.6 illustrates the relationship between the number of iterations and the number of files in this experiment. At the conclusion of each round, we captured the total time spent, from uploading the files to obtaining the extracted provenance information. In the first iteration, 1 file was uploaded to extract the provenance information, taking 49 seconds and costing \$0.14. During the second iteration, 2 files were used, which took 195 seconds and cost \$0.37. In the third iteration, 4 files were examined, taking 380 seconds and costing \$0.68. We continued with this approach, doubling the number of files 11 times (1024 files) and recording the time and cost, as illustrated in Table 6.4.

Iteration	Number of Files	Response Time	Cost (\$)
1	1	49 sec	0.14
2	2	195 sec	0.37
3	4	380 sec	0.68
4	8	548 sec	1.31
5	16	979 sec	2.57
6	32	1734 sec	4.77
7	64	3714 sec = 1.03 hr	10.30
8	128	7868 sec = 2.19 hr	19.27
9	265	18739 sec = 5.20 hr	39.10
10	512	43418 sec = 12.06 hr	79.89
11	1024	83181 sec = 23.1 hr	154.32

TABLE 6.4: Iteration results

Following the experiments, we began estimating the scalability of the extractor across all articles on exome sequencing experiments in PubMed (33597 files). This estimation was performed using several regression models. According to [Messaoud et al. \(2020, p.16\)](#), “regression analysis is a statistical method for modeling the relationships between the input variable and the continuous output variable”. To develop a prediction model, several common techniques can be used, including Linear Regression, Nonlinear Regression, and Logistic Regression ([Kott, 1991](#)).

Linear Regression is a statistical model used to summarise and study the relationship between two continuous values ([Kott, 1991](#)). The general form of this model is

$$y = mx + c$$

where c is the constant and m refers to the Regression Coefficient. There are several techniques that fall under this method, such as Least Squares or Ordinary Least Squares Regression.

Nonlinear Regression is another method for identifying a suitable nonlinear model to describe the relationship between one dependent variable and a set of independent variables. Its formula is

$$y = f(x, \beta) + \epsilon$$

where x represents the predictors and β refers to the nonlinear parameters. This type includes diverse techniques, such as the Gauss-Newton algorithm, Levenberg-Marquardt algorithm, Power laws and the Levenberg-Marquardt algorithm ([Dennis Jr et al., 1981](#)).

Logistic Regression is a predictive method for estimating or describing the variables associated with qualitative variables (often binary) from a group of explanatory variables, which can be quantitative or qualitative in nature ([Peng et al., 2002](#)). This type is very similar to linear regression in terms of practice ([Messaoud et al., 2020](#)). The formula for this model is

$$p(x) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}}$$

where x is the input variable and s represents the scale parameter.

To identify the most appropriate regression model to predict the estimated time from our provided results by representing the relationship between two variables: x and y where x is the number of files and y is the estimated response time, we conducted several steps.

Firstly, we determined the independent variable as the number of files, which is the input: x (1, 2, 4, 8, 16, 32, 64,...), and the dependent variable as the response time results from our experiment, which is the output: y (49, 195, 380, 548, 979, 1734,...).

Secondly, we determined the most common models to be examined, including the following regression analysis techniques: linear regression, polynomial regression, power-law regression and exponential regression. We used the `[scipy.optimize.curve_fit()]` function imported from `[scipy.optimize]` in Python, which utilises numerical optimization techniques to determine the optimal values for coefficients in each model. For example, the equation for linear regression is:

$$y = a \cdot x + b$$

and the fitted coefficients values using the above function, based on the observed data is $a = 82.1910$ and $b = -676.3674$. Using these coefficients, we can estimate the predicted values for scaling purposes.

Thirdly, we evaluated each regression model using metrics to determine the best model. We used a common metric to evaluate prediction models, the *Mean Absolute Error (MAE)* (Botchkarev, 2019). We calculated the *MAE* to determine the average absolute difference between the actual values and the predicted values using the following formula:

$$\text{MAE} = \left(\frac{1}{n}\right) \sum_{i=1}^n |y_i - x_i|$$

A smaller MAE indicates a better model. The MAE was calculated using a Python script with the `[mean_absolute_error]` function imported from `[sklearn.metrics]`.

Table 6.5 presents the equations of each model along with their respective MAE values. As mentioned earlier, the lowest MAE value indicates the best model. The comparison shows that the best-fitted model in this case is the power-law model, which achieved the lowest MAE value of 725.16, while the others exceeded 900. Therefore, we conclude that the power-law model is the most suitable for estimating the response time for all PubMed extractions.

Model	Generated Equation	MAE
Linear	$y = 82.1910 \cdot x - 676.3674$	908.6316726
Power Law	$y = 59.9557 \cdot x^{1.0452}$	725.1587329
Polynomial	$y = 0.000842 \cdot x^2 + 81.3906 \cdot x - 63.1726$	901.7246891
Exponential	$y = -5.2278 \times 10^{-17} \cdot e^{0.1 \cdot x}$	1.41×10^{27}

TABLE 6.5: Equations and MAE for time estimation.

Figure 6.7 compares the predicted results with the observed results, where the blue points represents the observed values and the orange line indicates the estimated values. Based on

the observed values compared to the predicted values using the power-law model, as shown in Table 6.6, we estimated the response time required for provenance extraction across all articles on exome sequencing experiments in PubMed (33,597 files). The estimated response time for 33,597 files is 3,226,411 seconds, equivalent to approximately 896.22 hours.

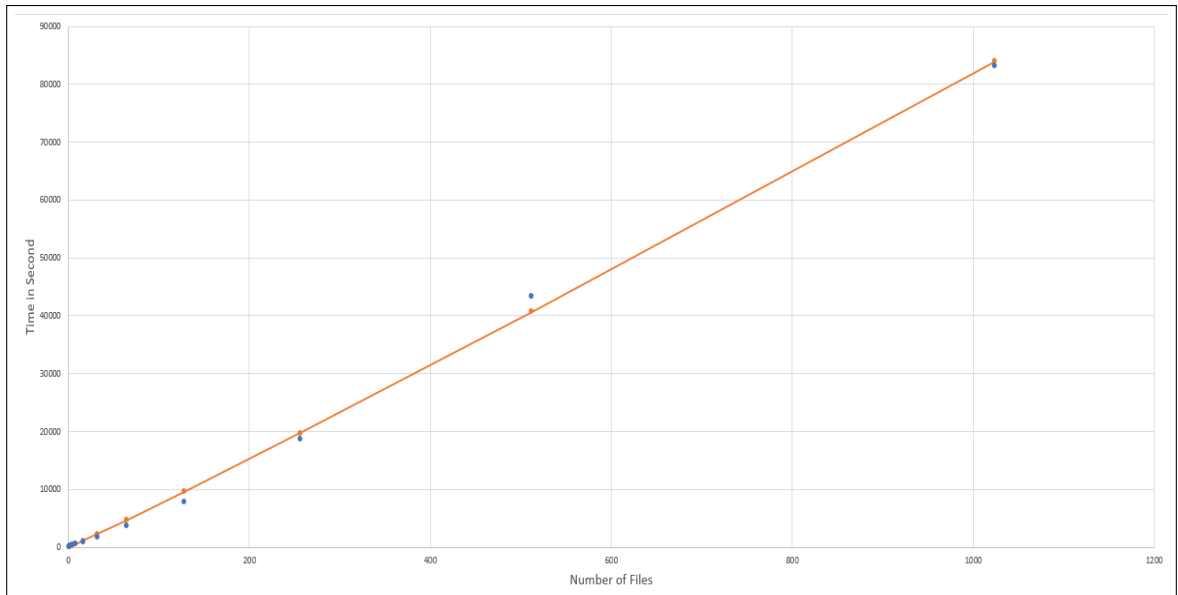


FIGURE 6.7: Estimated response time using the Power-Law model.

Number of Files	Observed Result in Seconds	Predicted Result in Seconds
1	49	59.9557
2	195	123.7277255
4	380	255.3310204
8	548	526.9144785
16	979	1087.36834
32	1734	2243.950308
64	3714	4630.733486
128	7868	9556.224374
256	18739	19720.72557
512	43418	40696.72308
1024	83181	83983.89115

TABLE 6.6: The observed results compared to the predicted results

We followed the same procedures to estimate the cost as well. To predict the estimated price from our results, we represented the relationship between two variables: x and y , where x is the number of files and y is the estimated cost.

The independent variable is the number of files, while the dependent variable is the cost, represented as y (e.g., 0.14, 0.37, 0.68, ...). Subsequently, the same regression analysis models were applied to determine the most appropriate model: linear regression, polynomial regression,

power-law regression and exponential regression. The coefficient values for each model were then identified using the same Python function. Finally, we calculated the MAE values, which showed two models performed well: Power Law (0.476) and Polynomial (0.474). Table 6.7 presents the equations and the MAE values.

Model	Generated Equation	MAE
Linear	$y = 0.1514 \cdot x^{0.1364}$	0.498
Power Law	$y = 0.1632 \cdot x^{0.9890}$	0.476
Polynomial	$y = -4.5923 \times 10^{-6} \cdot x^2 + 0.1557 \cdot x - 0.0927$	0.474
Exponential	$y = -5.7313 \times 10^{-17} \cdot e^{0.1 \cdot x}$	1.544×10^{27}

TABLE 6.7: Equations and MAE for cost estimation.

Figure 6.8 compares the observed results with the predicted results, where the blue points represent the observed values, and the orange line indicates the estimated values using the power-law model. Additionally, Figure 6.9 presents the observed and predicted results based on the polynomial model. Using the predicted values from the power-law and polynomial models, as shown in Table 6.8, we estimated the cost required for provenance extraction across all articles on exome sequencing experiments in PubMed. The estimated cost for processing 33,597 files is approximately \$4,889.10.

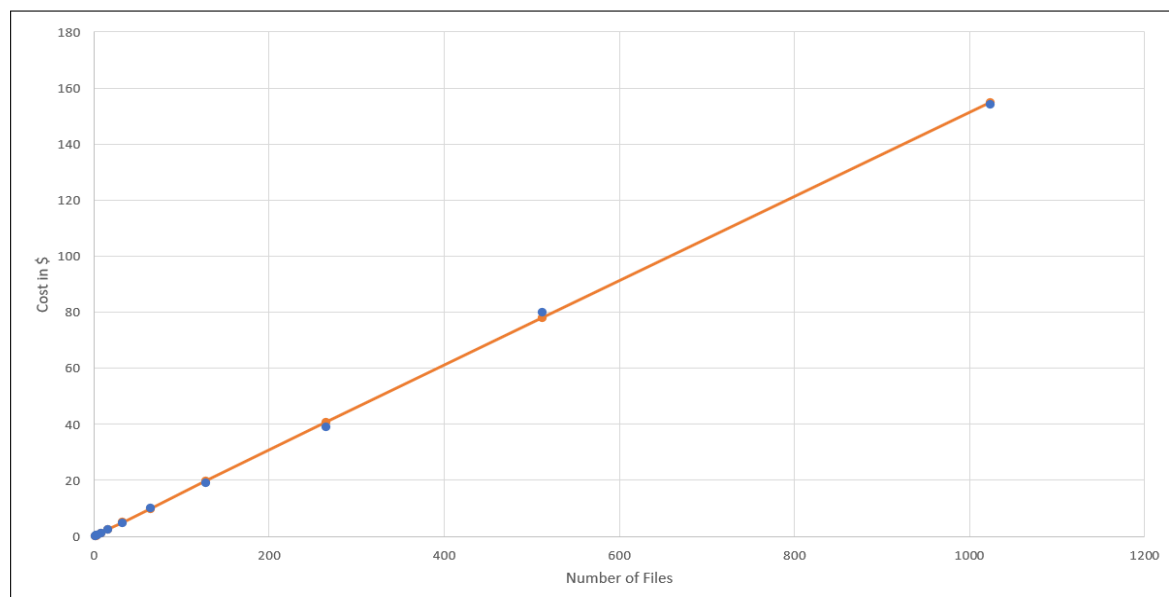


FIGURE 6.8: Cost estimation result using power-law equation.

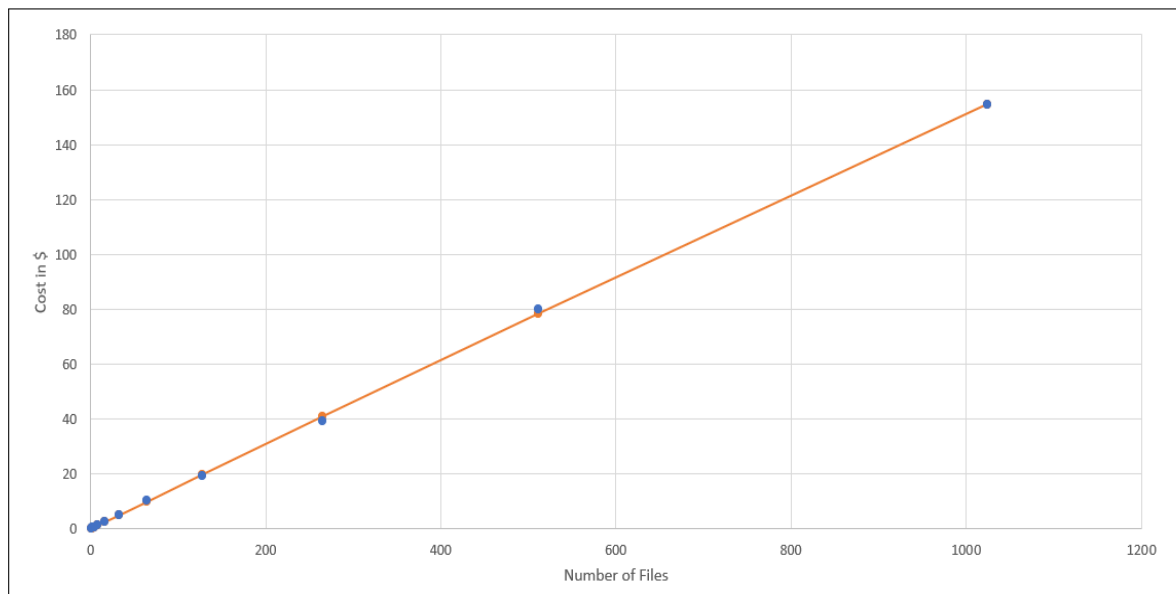


FIGURE 6.9: Cost estimation result using polynomial equation.

Number of Files	Observed Result (\$)	Power Law Prediction	Polynomial Prediction
1	0.14	0.1632	0.062995408
2	0.37	0.323920788	0.218681631
4	0.68	0.642920814	0.530026523
8	1.31	1.276074857	1.152606093
16	2.57	2.53276454	2.397324371
32	4.77	5.02705322	4.884997485
64	10.3	9.977739214	9.853289939
128	19.27	19.80390409	19.76165976
265	39.1	40.67338557	40.84530573
512	79.89	78.01680231	78.42185611
1024	154.32	154.8484318	154.5287244

TABLE 6.8: Comparison of observed results, power law predictions, and polynomial predictions.

6.4 Discussion

6.4.1 Prompt Patterns

As evident from the results of the intermediate prompts shown in Table 6.1, ChatGPT-4o effectively can access and extract the provenance information from scientific papers. Despite a notable variation in response accuracy — with all results promisingly above 65% — this variation can be attributed to some factors, including the complexity of some papers' formatting and the inherent limitations of session memory. The formatting complexity includes single or double column layouts, tables, figures and appendices. The limitations of session memory mean that each session has its own short-term memory.

The issue of formatting affecting text extraction from journal article PDFs is a common challenge in the biomedical text mining community ([Ramakrishnan et al., 2012](#)). PDF files often lack a consistent structure and variations in layout, font encoding and multi-column formatting can significantly hinder the reliable extraction of information at the sentence or section levels ([Gao et al., 2011](#)). In our extractor, the *fitz* library (PyMuPDF) was used to extract full-text content from PDFs. While *fitz* offers practical integration and speed, this library may affect the quality and consistency of the extracted information. Alternative approaches could involve the use of structured formats such as HTML or JATS XML, which can provide greater comprehensiveness and support a wider variety of formatting. Some publishers already expose structured formats, such as HTML, and offer more reliable access to content structure and metadata ([Comeau et al., 2019](#)). Although implementing and comparing such alternatives was beyond the scope of this research due to time constraints, this is a potential direction for future work.

Due to these formatting challenges, many classical biomedical text mining approaches have limited analysis to abstracts, which are readily available as plain text via some portals, such as PubMed. However, as discussed in Section 5.2.1.2, abstracts typically lack the detailed provenance information. This further highlights the need for robust methods capable of handling full-text content.

In the task of extracting provenance information using various prompt patterns, the results show the highest precision in patterns G (Scenario Pattern) and C (Question Refinement Pattern) for identifying all components, with the recall of these patterns exceeding 55%. In general, the model exhibited lower recall scores compared to precision, indicating to miss relevant provenance information. The highest recall scores, observed in prompt patterns D and H, were 0.66, indicating that they still exceed 60%.

However, prompt pattern H showed a high recall in relationships (0.681) and in the other components. Based on the results, we suggest improvements to the prompt patterns that may enhance the model's ability to retrieve relevant provenance information. In short, the complexity of the task, along with the models' lack of training for this task and the complexity of the format, are likely the main reasons for the reduced recall scores.

This claim aligns with [Ehsani et al. \(2025\)](#), which highlights that complex tasks, such as extraction tasks, are more challenging and less effective, even when using prompt engineering techniques. Crafting prompts for advanced or complex tasks may require comprehensive experience in prompt engineering techniques as well as a deep understanding of the tasks. The quality of developing effective prompts for complex tasks is challenging, particularly for non-expert users

(Zamfirescu-Pereira et al., 2023). It has been observed that the improvement of automatic prompt engineering can be affected by hallucination (Ye et al., 2023).

6.4.2 Scalability

In this scalability experiment, we observed a semi-linear trend in several models across different dataset sizes, suggesting that the extractor's performance remains consistent as the input increases. This result demonstrates that the extractor using GPT-4o can process larger datasets without significant changes in performance, including cost, which is an essential aspect of scalability. The cost primarily depends on factors: the type of model, (e.g., GPT-4o, a multimodal model), the pricing of input and output tokens, and the number of requests made to invoke the models. The response time can also depend on several factors: the number of requests, the API rate limits, which vary across models, and the usage tiers of the models, which can be upgraded based on the user's usage and expenditure on API services. For instance, Tier 3 allows users to use up to 800,000 tokens per minute but requires a minimum spending of \$5,000 per month.

Overall, our results confirm the capability of GPT-4o to extract information using designed prompts, which align with several studies (Bommasani et al., 2021; Jethani et al., 2023). Furthermore, we noticed that GPT-4o can handle and manage increased data volumes without encountering failures in invoking its API.

These findings highlight the importance of scalability for provenance extraction, particularly as biomedical datasets continue to grow online and require associated provenance to improve dataset search. Our results align with prior studies on scalability challenges in web applications (Sivakumar, 2024), particularly regarding cost and resource constraints.

However, it is important to optimise the extractor's performance while mitigating time and resource demands. These aspects are essential for optimising the performance of systems (Son and Kim, 2001). Potential strategies can be used for this optimisation, including refining prompts to enhance efficiency, upgrading critical hardware components (e.g., GPUs or memory), and streamlining the extraction process by removing unnecessary information, such as references, without preserving essential provenance information.

6.5 Summary

In this chapter, we presented the implementation of the extractor, utilising GPT-4. The architecture was explained in detail, encompassing its components, tools, and functions. Furthermore,

we discussed the experiment conducted to determine the most suitable prompt patterns for provenance extraction. Lastly, we examined the scalability experiment, which was designed to predict the extractor's response time and cost under scalable conditions. In the next chapter, we complement this empirical evaluation with a user experience study.

Chapter 7

Evaluating the Usefulness of Provenance Extraction for Dataset Search

This chapter presents the results of evaluating the performance of the extractor in terms of its completeness and correctness. The evaluation was conducted through a user experience study, using semi-structured interviews. In Section 7.2, we present the results of the qualitative analysis conducted on the data collected through the semi-structured interviews. In Section 7.3, we present the results of the quantitative analysis, using standard information retrieval metrics.

7.1 Demographics

As mentioned earlier in Section 3.5, the interviewees were recruited through the survey conducted in the earlier study, detailed in Section 3.3, where they were asked to provide their email addresses if they wished to participate. A total of 10 researchers participated in the interviews. All interviewees were biomedical researchers from the FoM at the UoS, working in various roles. Table 7.1 presents the demographic information of the participants.

Six participants were senior researchers holding diverse roles, including Research Fellow and Teaching Fellow, while the remaining participants were postgraduate researchers. All participants are specialised in different research domains, including clinical experimental science, cancer science, and genomics. Additionally, they worked with various types of omics data, as demonstrated by the papers they provided.

As stated in Section 3.5.1, we asked all participants to provide three papers that they had published based on an experiment, or papers that had been significant in their work. These papers were expected to include the workflow steps of conducting experiments that led to the generation of the data within the dataset.

Participants	Research Domains	Roles
P1	Genomics	Research Fellow
P2	Clinical and Experimental Science	Postgraduate Researcher
P3	Clinical and Experimental Science	Postgraduate Researcher
P4	Cancer Sciences	Research Fellow
P5	Genomics	Teaching Fellow
P6	Genomics	Postgraduate Researcher
P7	Clinical and Experimental Science	Research Fellow
P8	Cancer Sciences	Postgraduate Researcher
P9	Cancer Science	Research Fellow
P10	Genomics	Research Fellow

TABLE 7.1: Participants' research domains and roles

7.2 Qualitative Analysis Result

The aim of this study was to discover the quality of the extractor outcome cross different level experience of several biomedical researchers. In this part, all provided feedback was analysed qualitatively to evaluate the extraction results for the papers uploaded to the extractor, as detailed in section 3.5.

After completing the evaluation, the interviewees were asked to provide their opinions on the results concerning the four provenance quality dimensions, as defined in Section 3.5.1, and any other comments.

Overall, all the participants have expressed satisfaction about the provenance extraction. Several participants mentioned that the extractor performance was good (P6, P4). Additionally, the provenance information impressed several participants in terms of providing a comprehensive description and summary (P3, P5). A participant (P1) obtained all the main points of the provenance information in the provided papers. Other participants reported that the extractor can assist researchers in reading papers to gain specific details about datasets, potentially saving time by reducing the need to search for such details, including provenance information (P9, P1, P6).

An advantage of providing provenance information alongside datasets is its ability to assist in decision-making. Several participants emphasised that providing provenance information

helps them in the decision-making process and in selecting appropriate datasets (P5, P8). In addition, providing such provenance information will be sufficient to make a decision (P9, P7). A participant (P2) highlighted that establishing an association between actual datasets and their provenance information could significantly aid in the investigation and selection of suitable datasets. A participant (P1) mentioned that if this extractor were deployed and accessible, it would assist in providing the provenance of datasets from several papers, thereby facilitating the process of selecting an appropriate dataset for use. Table 7.2 includes several quotations related to the themes of performance and decision-making.

Theme	ID	Quotes
Overall performance	P6	"The extractor did a good job"
	P4	"In fact, I think nothing will ever be perfect, but this is already a pretty good run. "
	P3	"Yes, this is correct, and it's a good summary. "
	P5	"and also, a nice touch is the description, like the definition of the said technique. I really appreciate that."
	P1	"I was surprised it got all the main points, like the one thing with the minor detail — it's not even that important. "
	P9	"It's impressive. I mean, I think this is the next level. Hopefully, people won't have to read multiple papers at the same time to find specific details."
Decision-making	P5	"It's enough for me to make a decision"
	P8	"It's definitely valuable to download the dataset based on this information."
	P9	"I think that will definitely be enough to either say yes or look at it in more detail."
	P7	"I think it's enough to either make a decision or look into the paper just to check that what you want is there."
	P1	"If I had access to it, I would search for a certain thing, run all the papers through it, and then know if this is a good dataset to use "
	P2	"I could say yes, because it links actual datasets to the processes used to generate the data and to the researchers who created it. So, I think it is helpful for choosing a dataset to investigate."

TABLE 7.2: Participant quotations related to the themes of performance and decision-making.

Completeness is an interesting aspect highlighted by our participants during assessing the provenance information. Several participants mentioned that they gained sufficient information, indicating the completeness of the provenance information provided. Two participants (P9, P4) confirmed obtaining detailed and complete provenance information about the datasets. Another participant (P10) indicated that the completeness of the provenance information provided is sufficient for searching RNA sequencing datasets related to a rare disease. Participant (P7) gained complete information regarding the aim, an understanding of the experimentation performed, and details about how the data was manipulated. Another participant (P4) stated that the provenance information was complete and assisted in evaluating the dataset. They also mentioned that it

provided details about how much data was produced by their group and how much was used from other sources.

Regarding the missing information in the extracted provenance, several participants mentioned that no important information was missing or needed to be included (P3, P4). Two participants expressed that they expected missing information, however the provenance was sufficient (P9, P4). A participant expressed that while it is hard to achieve absolute perfection, this is already a great effort (P10). However, the extent of missing information varied among the three papers provided to the extractor (P2). Additionally, a participant referred to missing information in the relationship components, which makes the provenance information difficult to understand (P4).

We defined completeness for our participants as how much provenance is missing or exceeds the actual amount of collectible provenance, known as “overcomplete”. This definition was taken from [Cheah and Plale \(2014\)](#). Nine participants confirmed that the completeness of the provided information ranged between excellent and good, while the last participant described it as fair. Table 7.3 includes several quotations related to the themes of provenance completeness and missing information.

Theme	ID	Quotes
Provenance completeness	P9	“Obviously, this was really good because it covered a lot of details, and completeness.”
	P10	“I think it depends on how specific you’re looking for. If you’re looking for a specific dataset with very precise terms, you might want to go in and go further. But if I’m looking for, I don’t know, for example, blood RNA sequencing from a rare disease, then that’s enough.”
	P7	“It’s clear how easy it is to pick out the key information here. You now get a clear idea of the main points, like the aims, and an understanding of what data manipulation or experimentation was performed.”
	P4	“This makes it very easy to pick out what is necessary to evaluate whether you want to proceed with the paper or not. What’s also quite nice is that you can tell how much of the data comes from the group and how much comes from outside the group, which is quite useful. Overall, I think it’s pretty cool.”
Missing information	P2	“I think the last one has less information than the other two, is less comprehensive”
	P3	“There is no missing information that should be included here.”
	P9	“It’s like half correct and a bit missing, it’s not fair to say that it’s all wrong. I think it’s like both, if that makes sense.”
	P4	“So that makes it a little harder to decipher what the relationships are.” “I think nothing will ever be perfect, but this is already a pretty good run.”

TABLE 7.3: Participant quotations related to the themes of provenance completeness and missing information.

Accuracy and correctness stood out as crucial parts highlighted by the participants in this study. All participants were asked to share their opinions about the accuracy and correctness of the provided provenance information. According to [Black and van Nederpelt \(2020\)](#), the

correctness aspect includes several attributes, including accuracy, consistency, unambiguity, and homogeneity of the provided information. Several participants asserted that the correctness of the extracted provenance was excellent (P10, P9, P4), while all other participants referred to it as having good accuracy. Table 7.4 includes several quotations related to provenance accuracy.

Theme	ID	Quotes
Provenance accuracy and correctness	P10	"I thought it was very accurate."
	P4	"I think the information is almost always accurate, which is quite nice. The one thing I appreciate is that this is just an extracted form, and it has been prepared to meet our needs."

TABLE 7.4: Quotes of participants related to provenance accuracy.

Although all the above advantages were noted, several participants observed the integration of some provenance information into a single component. Participant (P2) mentioned that, despite the provenance information being well-classified, it combined two processes under one activity. For example, one participant (P8) observed that a researcher (agent) from the analysis team was added to another team (the research investigation team), which was expected to be identified as a separate component. 7.5 includes several quotations regarding this disadvantage.

Theme	ID	Quotes
Provenance component integration	P3	"However, I would expect to see that as a separate entity"
	P2	"I did say that it was quite good at sorting the information. I did notice that sometimes it will pair two processes under one thing. So I think it would make sense to have this as two separate activities."
	P8	"This one has conducted the analysis and then merged it with the research investigators and their roles in the study, whereas I would expect to see that as a separate entity."

TABLE 7.5: Quotes of participants related to component integration.

7.3 Quantitative Analysis Result

As mentioned earlier in Section 3.5.3, we asked participants to evaluate the quality of the extracted provenance information by highlighting the correct and incorrect components. For clarification, when a paper is uploaded to the extractor, it produces a set of PROV-DM components, denoted as $LLM(P) = e, a, c, r, u, o$, where $LLM(P)$ refers to the provenance extracted using the LLM, and e, a, c, r, u, o represent the entities, activities, agents or relationships extracted by the model. The annotated evaluation is then performed by an expert (Pe), who identifies components as either correct or incorrect, yielding the set $ee, aa, cw, rr, uu, io, n$. Here, the correct components are ee, aa, rr, uu , while the incorrect ones are cw, io . In this example, precision is calculated as

the proportion of relevant (i.e. correct) provenance components retrieved, divided by the total number of components retrieved. This is expressed as:

$$\text{Precision} = \frac{|\{e, a, r, u\}|}{|\{e, a, c, r, u, o\}|} = \frac{4}{6} \approx 0.66$$

The highest average precision of all components (Entities, Activities, Agents, Relationships) was 1, achieved for 10 papers provided by eight participants. Thirteen papers achieved the second highest average precision, ranging between 0.98 and 0.90, provided by seven participants. Subsequently, four papers scored average precision between 0.87 and 0.84. The lowest calculated average precision ranged between 0.71 and 0.77, observed for only three papers. Table 7.6 presents the precision of each component and the average values across all components. Overall, the average value across all components is approximately 0.905, indicating good results.

ID	Paper	Precision for Entities	Precision for Activities	Precision for Agents	Precision for Relationships	Precision for All Components
1	P1	0.85	1	1	1	0.96
	P2	0.75	1	1	1	0.93
	P3	1	1	0.66	1	0.91
2	P1	1	1	1	1	1
	P2	0.76	1	0.75	0.93	0.86
	P3	1	1	1	0.91	0.97
3	P1	0.81	0.55	1	1	0.84
	P2	1	1	1	0.63	0.90
	P3	1	1	1	0.90	0.97
4	P1	0.94	1	1	1	0.98
	P2	1	1	1	1	1
	P3	0.87	0.85	1	1	0.93
5	P1	0.97	0.96	1	1	0.98
	P2	1	1	1	1	1
	P3	0.8333	1	1	1	0.95
6	P1	1	1	1	1	1
	P2	1	1	1	0.85	0.96
	P3	0.92	0.93	1	0.89	0.93
7	P1	0.80	0.8	1	0.9	0.87
	P2	1	1	1	1	1
	P3	0.92	0.81	0.40	0.76	0.7
8	P1	1	1	1	1	1
	P2	1	1	1	0.5	0.87
	P3	1	1	1	1	1
9	P1	0.88	1	0.5	0.84	0.77
	P2	1	1	1	1	1
	P3	1	1	1	1	1
10	P1	0.87	0.85	0.66	0.81	0.79
	P2	0.81	0.952	1	1	0.92
	P3	1	1	1	1	1

TABLE 7.6: Precision for provenance information components across biomedical papers

For recall measurement, it is calculated as the proportion of relevant provenance information retrieved to all relevant components. This is expressed as $\{e, a, r, u\} / \{e, a, r, u\}$, which equals $4/4 = 1$. The highest average recall obtained was 1, achieved for 27 papers provided by all participants. All components retrieved by the extractor were highlighted as relevant or correct by the participants. The remaining average recall values for the three papers ranged between 0.85 and 0.95, indicating excellent results. The overall average value for all components across the papers was 0.991. Table 7.7 exhibits the recall of each component and the average values across all components for each paper.

ID	Paper	Recall for Entities	Recall for Activities	Recall for Agents	Recall for Relationships	Recall for All Components
1	P1	1	1	1	1	1
	P2	1	1	1	1	1
	P3	1	1	1	1	1
2	P1	1	1	1	1	1
	P2	1	1	0.75	1	0.93
	P3	1	1	1	1	1
3	P1	1	1	1	1	1
	P2	1	1	1	1	1
	P3	1	1	1	1	1
4	P1	1	1	1	1	1
	P2	1	1	1	1	1
	P3	1	1	1	1	1
5	P1	1	1	1	1	1
	P2	1	1	1	1	1
	P3	1	1	1	1	1
6	P1	1	1	1	1	1
	P2	1	1	1	1	1
	P3	1	1	1	1	1
7	P1	1	1	1	1	1
	P2	1	1	1	1	1
	P3	1	1	1	1	1
8	P1	1	1	1	1	1
	P2	1	1	1	1	1
	P3	1	1	1	1	1
9	P1	1	1	1	1	1
	P2	1	1	1	1	1
	P3	1	1	1	1	1
10	P1	1	1	1	1	1
	P2	1	0.869	1	1	0.95
	P3	0.571	1	1	0.83	0.85

TABLE 7.7: Recall for provenance information components across biomedical papers

Although the components were correct in several cases, we observed that their descriptions were either incorrect, partially missing, or repeated. For example, BALB/c mice (E1) was classified as an entity and provided with the following description: “Type of mice used in the experiments. - Obtained from Charles River Laboratories and maintained in local facilities.” However, Participant (P4) noted that this was incorrect, as the mice were not maintained in local

facilities. In another example, Participant (P5) confirmed that the following activity was correct, but its description was insufficient to fully understand the concept of the activity: “Systematic Reanalysis (A1) - Description: Structured reanalysis of ES data from previously undiagnosed patients. - Performed in three steps, including variant reevaluation and resequencing.”

Therefore, we re-evaluated the quality and accuracy of the extracted results, considering the entire component incorrect if it contained missing, incorrect, or repeated information. From our perspective, referring to a component as correct while ignoring the quality of its description can be deceptive and lead to a misunderstanding. Providing incorrect provenance information or an incomplete description of the provenance could ultimately impact dataset selection and search.

The average precision results showed modest variation. Only seven papers achieved a precision score of 1, while 11 papers gained the second highest average precision, ranging between 0.98 and 0.91. Thereafter, eight papers fell within the range of 0.87 to 0.80, representing the third highest precision average. The remaining papers displayed the lowest results, with precision values ranging between 0.78 and 0.61. Overall, the average precision value for all component across all paper was slight decreased to 0.901. Table 7.8 displayed the precision of each component and the average values across all components.

Regarding recall average, there was a noticeable variation in the average compared to the previous recall result, with the overall value decreasing from 0.991 to 0.952. As illustrated in Table 7.9, recall values generally decreased across several papers for different components. The number of papers with the highest average recall of 1 dropped from 27 to 16. Ten papers scored between 0.99 and 0.90, while the remaining four papers were within the range of 0.82 to 0.71. However, this result still demonstrates a high level of accuracy, reflecting the positive satisfaction with the extracted provenance information.

7.4 Discussion

Our results demonstrate that the extractor produces high-quality outcomes, and the participants’ feedback was positive, highlighting the potential of LLMs for provenance extraction. Participants confirmed that providing provenance information alongside datasets enhances their decision-making process and facilitates the selection of appropriate datasets, thereby contributing to advancements in the dataset search domain. The extractor demonstrates significant capabilities in terms of completeness, accuracy and correctness, as evidenced by our measurements of precision and recall.

ID	Paper	Precision for Entities	Precision for Activi- ties	Precision for Agents	Precision for Rela- tionships	Precision for All Compo- nents
1	P1	0.857	1	1	1	0.96
	P2	0.75	1	1	0.96	0.95
	P3	1	1	0.66	1	0.91
2	P1	1	1	1	1	1
	P2	0.76	1	0.75	0.93	0.86
	P3	0.93	0.85	1	0.82	0.90
3	P1	0.81	0.56	1	1	0.84
	P2	1	0.89	0.89	0.63	0.84
	P3	1	1	1	0.90	.97
4	P1	0.83	0.76	0.67	1	0.81
	P2	1	0.53	0.60	1	0.78
	P3	0.87	0.85	0.67	0.94	0.83
5	P1	0.97	0.96	1	1	0.98
	P2	1	1	1	1	1
	P3	0.83	1	1	1	0.95
6	P1	1	1	1	1	1
	P2	1	1	1	0.85	0.96
	P3	0.92	0.93	1	0.89	0.93
7	P1	0.8	0.8	1	0.90	0.87
	P2	1	1	1	1	1
	P3	0.92	0.81	0.40	0.76	0.72
8	P1	1	1	1	1	1
	P2	1	1	1	0.5	0.87
	P3	1	1	1	1	1
9	P1	0.88	0.87	0.5	0.84	0.77
	P2	0.90	1	1	1	0.97
	P3	1	1	1	1	1
10	P1	0.87	0.85	0.67	0.81	0.80
	P2	0.67	0.71	0.67	0.40	0.61
	P3	1	1	1	0.83	0.95

TABLE 7.8: Precision for description of provenance information components across biomedical papers.

Despite these promising results, several challenges remain as areas for improvement. The performance of LLMs can be influenced by several factors such as the paper’s length, its density of information and its formatting. In our experiment, some papers exceeded 30 pages, often featuring two columns and appendices. These characteristics can affect token and context length constraints, leading to incomplete processing or potential inaccuracies in the extracted information.

While advancements have been made to address context length constraints, these limitations continue to impact the performance of LLMs (Kamath et al., 2024). Moreover, the explainability of LLMs is often affected by their sheer complexity and size, as many models include billions of parameters, complicating the interpretation of their decision-making processes (Hada and Shevade, 2021; Gao et al., 2023). Another critical issue is hallucination, where LLMs generate

ID	Paper	Recall for Entities	Recall for Activities	Recall for Agents	Recall for Relationships	Recall for All Components
1	P1	1	1	1	1	1
	P2	1	1	1	0.96	0.99
	P3	1	1	1	1	1
2	P1	1	1	1	1	1
	P2	1	1	0.75	1	0.93
	P3	0.93	0.92	1	0.90	0.93
3	P1	1	1	1	1	1
	P2	1	0.89	1	1	0.97
	P3	1	1	1	1	1
4	P1	0.88	0.76	0.67	1	0.82
	P2	1	0.53	0.6	1	0.78
	P3	1	1	0.67	0.94	0.90
5	P1	0.97	0.92	1	1	0.97
	P2	0.93	1	1	1	0.98
	P3	1	1	1	1	1
6	P1	1	1	1	1	1
	P2	1	1	1	1	1
	P3	1	1	1	1	1
7	P1	1	1	1	1	1
	P2	1	1	1	1	1
	P3	1	1	1	1	1
8	P1	1	0.67	1	1	0.91
	P2	1	1	1	1	1
	P3	1	1	1	1	1
9	P1	1	0.87	1	1	0.96
	P2	0.90	1	1	0.83	0.95
	P3	1	1	1	1	1
10	P1	1	1	1	1	1
	P2	0.81	0.65	0.67	0.72	0.71
	P3	0.57	0.67	1	0.83	0.76

TABLE 7.9: Recall for description of provenance information components across biomedical papers.

misleading or inaccurate information in an attempt to provide knowledge or context to address gaps (Liu et al., 2024a). This issue can lead to users receiving misleading or incorrect answers to their queries.

Several studies (Lewis, 2019; Abedu et al., 2024; Chen et al., 2024) highlight that LLMs face inherent token limitations, requiring documents to be split into chunks to overcome this constraint. This challenge can result in the generation of inaccurate responses or disrupt the document’s coherence (Wang et al., 2024). In conclusion, incorrect tokenization for lengthy texts can hinder the understanding of input, resulting in unsatisfactory output.

7.5 Summary

In this study, we aimed to assess the usefulness of provenance extraction for biomedical dataset search through a user experience study. Our findings demonstrate the considerable value of providing provenance information, as highlighted by our participants — ten biomedical researchers. These results reflect the ability of our extractor to infer provenance information from biomedical papers, with participants noting that the extracted provenance was of high quality.

In this chapter, we discussed the results of the user experience evaluation. This evaluation was conducted using both qualitative and quantitative analyses, employing various metrics, including completeness, correctness and accuracy.

Chapter 8

Discussion

This chapter summarises the main findings from Chapters 4–7, linking them to broader literature in IR, text mining, LLMs, visualisation and biomedical research practices. While each results chapter includes its own discussion, the focus here is on synthesising themes and identifying implications.

The studies presented in this thesis explore the enhancement of biomedical dataset search through the integration of provenance metadata in visualisations, drawing on insights and requirements from biomedical researchers' search behaviours, workflow visualisation practices, prompt engineering for provenance extraction and evaluations of presentation methods. The findings of these studies demonstrate that the provenance information of experimental data within datasets is not only additional metadata but a key element that strengthens dataset discovery, evaluation and reuse in biomedical domains.

From a biomedical research practices perspective, the first study (Chapter 4) revealed that provenance metadata addresses specific needs of biomedical researchers, particularly supporting experimental reproducibility and enabling the reuse of datasets. This study outlined three current strategies for searching datasets starting with search in public sources, followed by literature search and finally using social networks. In addition, this study revealed that biomedical researchers face distinctive challenges, including insufficient metadata quality and a lack of detailed provenance information. These findings are consistent with earlier research (Buneman and Tan, 2007; Collins and Tabak, 2014; Baum et al., 2017; Liu et al., 2020; Gregory et al., 2020), which emphasised the need for improved dataset metadata and highlighted the scarcity of provenance information associated with datasets. Provenance information can play a crucial role in ensuring the reproducibility and trustworthiness of research results (Valdez et al., 2017; Gierend et al., 2024). This study extends prior work by specifying biomedical-specific

requirements, such as integrating public source and literature searches into a unified process and leveraging social networks for dataset sharing. Given the importance of obtaining detailed provenance information for datasets, this integration of strategies does not guarantee the retrieval of sufficient or accurate provenance information, nor does it ensure a timely response — factors that may negatively affect both the research process and its outcomes.

While several data journals facilitate traditional scholarly communication to support biomedical data sharing and reuse (Fan et al., 2013), various social practices, including ethical, privacy, and acknowledgment considerations, may indirectly affect data exchange and reuse (Federer et al., 2015). Furthermore, sharing a dataset without sufficient metadata or provenance information can affect its usefulness; therefore, it has been recommended that standards for metadata and provenance information be improved to better support such tasks (Federer et al., 2015). Although scholarly communication is crucial across various domains, it faces significant challenges, particularly its lengthy process of publishing data within datasets. In the context of biomedical research, such delays can hinder timely dataset availability, limit opportunities for data reuse, and prevent the provision of detailed provenance information. This, in turn, may impact researchers' ability to assess dataset and trace research workflows — core requirements for effective dataset discovery and reuse.

From an IR perspective, the current search strategies for searching datasets, as observed in Chapter 4, which involve sequentially combining public source, literature and social network search, parallel query expansion in IR, where diverse information sources refine the researcher's understanding to improve dataset selection and reuse. The application of these strategies may depend on the researcher's skills and experience, including the ability to design effective prompts for each strategy and to integrate heterogeneous datasets from multiple sources. There is scope for future dataset search systems to automate these strategies. Techniques such as semantic search and LLMs could be employed to reduce reliance on manual integration and to retrieve high-quality datasets across varying levels of user expertise.

In the context of LLMs and prompt engineering, the evaluation in Chapter 6 showed that Scenario and Question Refinement patterns yielded the highest precision and recall for provenance extraction from biomedical literature. While consistent with emerging research on prompt engineering for complex extraction tasks (Ehsani et al., 2025), the observed recall limitations echo concerns about token limits and hallucination in LLM outputs (Liu et al., 2024a; Ye et al., 2023). These findings highlight the importance of post-extraction validation and optimisation to ensure reliability of results. LLMs are continually expanding, with ongoing improvements in their capabilities to meet users' needs. However, to be valuable for the biomedical community in the context of dataset search and reuse, the most critical challenge to address is the hallucination

problem. This community depends on scientifically accurate information, and given the time constraints researchers face, providing a tool that produces inaccurate results due to hallucinations would be more harmful than requiring them to manually review datasets, as is currently the practice. Future research on mitigating hallucinations in LLMs, such as that described by (Zhang et al., 2025), could further enhance the reliability and applicability of LLMs.

The role of formatting complexity in extracting provenance information from PDFs was also identified as a key issue in this study. These challenges of extracting information from PDFs align with long-standing issues in biomedical text mining (Cohen and Hersh, 2005; Bhargava et al., 2017). Many classical approaches have focused on abstracts, owing to their more reliable accessibility via portals such as PubMed. However, this study highlights the necessity of robust full-text extraction to meet provenance requirements. Thus, adopting structured publishing standards (e.g., JATS XML) for full-text articles can substantially enhance provenance extraction efforts.

Several prior studies on workflow visual representations, visualisations, and graphical abstracts have aimed to present research procedures in a way that is easy to understand (Von Rosing et al., 2015; Willoughby and Frey, 2017; Wratten et al., 2021). Building on this idea, we designed an activity-centred diagram to present the provenance of data within datasets. We then assessed the effectiveness of this visual representation — which incorporates provenance information from biomedical experimental data contained in public datasets — in Chapter 5. The results of this assessment demonstrated that provenance metadata within workflow visualisations significantly enhances biomedical dataset search and reuse, outperforming other types of metadata used in platforms such as DataMed and PubMed. In addition, Chapter 7 confirmed that presenting provenance information generated by LLMs enhances user understanding and supports informed decision-making in dataset search. This representation of provenance metadata can improve the process of understanding and evaluating datasets, thereby supporting the application of the FAIR principles in dataset search. Furthermore, our approach aligns with Sahoo et al. (2023), who demonstrate provenance’s potential to reduce uncertainty and improve decision-making. Although these visualisations were created manually for the purposes of this study, they align with the broader literature on workflow visualisation in biomedicine and the use of visual or graphical abstracts. However, creating such visualisations manually is time-consuming; therefore, automating the generation of these graphs represents an important step towards improving dataset search. Future research on automated visualisation generation using LLMs, such as that described by Khan et al. (2025), could further advance the automation of visual representations, workflow visualisations, and graphical abstracts.

The results of the studies in this research indicate that the integration of provenance metadata into biomedical dataset search interfaces offers benefits to the process of decision-making. Additionally, the studies support the FAIR principles by enabling more transparent dataset evaluation, reducing uncertainty about dataset suitability for users' needs, and fostering interdisciplinary trust. However, the work also identifies persistent barriers: the absence of automated provenance graph generation and limited publisher adoption of structured full-text formats (JATS XML).

8.1 Summary

This chapter presents an overall discussion of how provenance metadata can enhance biomedical dataset search — from identifying community-specific requirements to developing extraction and visualisation approaches. By linking these contributions to the broader literature in IR, text mining, LLMs, visualisation and biomedical research practices, it establishes a foundation for both theoretical and applied advancements. The findings highlight the critical role of provenance in improving dataset discovery and reusability, with implications that extend beyond the biomedical domain.

Chapter 9

Conclusions

In this thesis, we investigated the issue of dataset search in the biomedical research domain and explored ways to enhance the search process for public datasets. Our main goal was to assist biomedical researchers in the dataset search process, thereby enhancing the findability and reusability of datasets. We began by examining the field of dataset search, including existing tools, systems, techniques and benchmarks, and subsequently developed a corresponding taxonomy (Almuntashiri et al., 2022). To answer RQ1 and gain insights into biomedical researchers' dataset search behaviours, challenges, and requirements, we conducted semi-structured interviews. Our findings revealed several issues, including insufficient metadata, missing provenance information, concerns about data quality and limitations in accessibility. These challenges informed our decision to focus on providing provenance information to enhance the dataset search process and better support researchers' needs.

To address RQ2, we then designed two imaginary dataset search tasks, each providing multiple options, with one incorporating provenance information extracted from biomedical publications, and distributed them via a questionnaire to biomedical researchers. The aim of this study was to assess the impact of providing provenance information when searching for biomedical datasets and to evaluate how this information could aid researchers in searching for suitable datasets, i.e., they identify the provenance information elements needed when searching for datasets.

After evaluating the effectiveness of provenance information in biomedical dataset search, we developed an automated extraction tool aimed at extracting/infering provenance information from biomedical publications, addressing RQ3. This tool can assist researchers in extracting valuable provenance information buried within publications, thereby improving the dataset search process, and enhancing findability and reusability of datasets. In this tool, we utilised an LLM to extract/infer provenance information, as it has demonstrated its ability in information

extraction tasks. Additionally, we developed several specialised prompts to optimise the LLM's performance and conducted experiments to identify the most effective one for use. However, this technique still requires further improvements.

Finally, we assessed the tool's performance through a user experience study, which allowed us to answer RQ4. We asked ten biomedical researchers, who had published articles based on experimental data within public datasets, to upload three papers into the tool and evaluate the usefulness of the extracted provenance. This evaluation focus on the completeness and correctness of the provenance information.

Through this research, we found several requirements that need to be addressed to improve datasets, which can be summarised as follows:

- The dataset search process in biomedicine is essential, aiming to assist biomedical researchers in finding and reusing existing datasets. However, researchers face several challenges with current approaches to dataset search, including a lack of metadata that adequately describe public datasets. Our findings indicate that researchers have specific requirements for improving the effectiveness of biomedical dataset search.
- We find that providing provenance metadata has a positive influence on dataset search. Provenance information enhances the ability of biomedical researchers to assess public datasets more effectively than the current metadata available in platforms such as DataMed and PubMed. It provides a deeper understanding of the history of the data within datasets and facilitates their reuse, which is particularly valuable for reproducibility tasks. In addition, provenance information helps reduce uncertainty when assessing online datasets, thereby supporting the implementation of the FAIR principles in dataset search.
- We demonstrate the ability of LLMs, specifically ChatGPT-4o, to infer provenance information from publications based on datasets, achieving high-quality results through the use of enhanced prompts. In addition, the model demonstrates scalability in inferring provenance from all publications related to exome sequencing experiments in PubMed. Furthermore, LLMs can convert the inferred provenance from unstructured text into the PROV-DM format.
- We find that the provenance information inferred from publications using ChatGPT-4o was well received by biomedical researchers in a user experience study, assisting them in understanding the history of the data and supporting its find and reuse.

9.1 Future Work

The dataset search domain has gained significant attention over the past decade from various communities, including the information retrieval, open data, and database communities. In addition, this domain has attracted increasing interest in several international scientific events, such as the 'DATA: SEARCH'18' workshop. At the same time, dataset search has been increasingly utilised across multiple scientific domains to facilitate dataset discovery and reuse, supporting research tasks such as reproducibility. Given the vast number of public datasets available in open portals and domain-specific repositories, we believe there remains a significant need to explore the unique characteristics of each data type, as different scientific communities prioritise their specific needs and research requirements.

Regarding dataset search in biomedical research domains, which is the focus of this thesis, there are multiple research directions to consider for future work. One research direction is to improve metadata quality in public sources to better meet the needs of dataset searchers. As we found in Chapter 4, the quality of metadata in public repositories influences the dataset search process. This finding aligns with prior research ([Chapman et al., 2020](#); [Löffler et al., 2021](#)). In our study, we addressed the lack of provenance information in metadata within biomedical dataset search, which can support decision-making and the reusability of existing datasets.

However, providing provenance information is not only essential in this context; it is also critical across various scientific domains and tasks. As confirmed by [Wittner et al. \(2023\)](#), tracking and providing provenance to end users can enhance the reproducibility of research outcomes, increase trust in experimental results for reuse, and support the assessment of data quality. Nevertheless, a gap remains in the automatic or inferred capture of provenance metadata. Addressing this gap is crucial for strengthening the trustworthiness and reusability of scientific datasets.

Along with provenance metadata, there remains a gap in enhancing metadata with respect to findability and accessibility, as identified in our study (see section 4.2.6). As confirmed by [Löffler et al. \(2021\)](#); [Ding et al. \(2014\)](#), effective dataset retrieval depends on the availability of metadata that describes the methods for finding and accessing datasets—both of which are key components of the FAIR data principles. Moreover, such metadata should be comprehensive, encompassing all relevant elements necessary to understand the history of the data, as well as how to find and access the data for reuse. Thus, an open research direction lies in developing methods to capture accessibility information about datasets and integrate this information with provenance metadata, thereby strengthening adherence to the FAIR data principles.

Another research direction is to investigate and develop an automated method suitable for presenting all provenance information in accordance with the provenance standard mentioned above. Although previous studies, such as [Deutch et al. \(2015\)](#), have focused on the presentation of provenance information, a key gap remains in the automatic interpretation and adaptation of provenance information for multiple types of consumers, ranging from experts to non-experts. [Wittner et al. \(2023\)](#) highlight that provenance information spans the entire research lifecycle, from material acquisition to the derivation of results. Therefore, a comprehensive method should not only capture or infer the full range of provenance information but also automatically tailor its presentation to suit the needs and levels of understanding of diverse researchers, thereby facilitating the reading, interpretation, and reuse of the data.

We hope that our research will contribute to assisting biomedical researchers in searching for public datasets and enhancing the reusability and reproducibility of experimental datasets. In addition, we aspire for other researchers to further advance the development of dataset search, both in general and within the biomedical research domain.

Appendix A

Human Study 1

Participant Information Sheet

Study Title: Dataset search for ML pipelines

Researcher: Abdullah Almontashiri
ERGO number: 73032

You are being invited to take part in the above research study. To help you decide whether you would like to take part or not, it is important that you understand why the research is being done and what it will involve. Please read the information below carefully and ask questions if anything is not clear or you would like more information before you decide to take part in this research. You may like to discuss it with others but it is up to you to decide whether or not to take part. If you are happy to participate you will be asked to sign a consent form.

What is the research about?

There are already several existing services and methods such as Google Dataset Search that assist users in finding datasets; they include several new features to meet users' needs of datasets. Nevertheless, there is a lack of the methods of expressing requirements of datasets to be used for machine learning (ML) methods and pipelines beyond searching for content and format. Due to the large number of datasets that exist on the web, researchers/ scientists can encounter a number of difficulties when trying to discover appropriate datasets for ML. Choices of dataset spread can impact the performance of a machine learning system. The dataset search field is novel, and to date, there has been a lack of research on dataset search. The purpose of this research study is to supplement such efforts and develop new methods and tools or improve existing dataset search methods that could help find datasets for use in ML pipelines. It will explore what types of additional information ML developers need, how to express this in a query, how to measure it over a dataset and return results.

Why have I been asked to participate?

The primary goal of this research is to discover the dataset search requirements for machine learning models in order to design a new approach. Therefore, you have been invited to participate since your knowledge of the field is sufficient and you are capable of answering the questions.

What will happen to me if I take part?

By conducting the interview, you will help us in the dataset search domain. Your answers will help to find out the requirements of dataset search for ML models. Before conducting this interview, you need to read the consent form and then decide to agree or disagree. If your decision is to agree, you will start to answer some questions based on your experience in dataset search for ML purposes.

Are there any benefits in my taking part?

There is no benefit that will be given to participants in this study. The participation in this study will be voluntary. The participants will contribute to develop the domain of dataset search for ML models. This contribution can improve the development of dataset search as well as facilitate the process of discovering datasets for ML purposes.

Are there any risks involved?

No. This study was approved by the Ethics committee at the University. There are no expected risks identified.

What data will be collected?

The aim of this study is to collect data to answer only research questions. Therefore, personal information will not be collected that could determine the identity of the participants. Regarding demographics data, the level of experience in searching dataset for ML purposes can be collected.

Despite there is no personal information will be collected, all collected data will be anonymous and confidential with Data Protection Laws.

I would like to record the interviews and then transcript them. Therefore, the participants will be informed that I intend to record the interview and this is optional; they have the right to agree or disagree. A tick box for the audio recording agreement will be provided in the consent form.

Will my participation be confidential?

Your participation and the information we collect about you during the course of the research will be kept strictly confidential.

Only members of the research team and responsible members of the University of Southampton may be given access to data about you for monitoring purposes and/or to carry out an audit of the study to ensure that the research is complying with applicable regulations. Individuals from regulatory authorities (people who check that we are carrying out the study correctly) may require access to your data. All of these people have a duty to keep your information, as a research participant, strictly confidential.

According to university policy, the collected data will be stored for ten years on the University server. All files will be encrypted to be more confidential. I intend to quote directly from the answers of the participant, if I need a direct quotation. Any audio records will be destroyed after the transcribing process is done. Then, the transcribed data will be stored with all the files of this study on the University server.

Do I have to take part?

No, it is entirely up to you to decide whether or not to take part. If you decide you want to take part, you will need to sign a consent form to show you have agreed to take part.

The participants, who will conduct the interview, will be asked to sign the consent form before starting the conversation.

What happens if I change my mind?

You have the right to change your mind and withdraw at any time without giving a reason and without your participant rights being affected.

All collected data in this interview will be discarded and deleted.

What will happen to the results of the research?

Your personal details will remain strictly confidential. Research findings made available in any reports or publications will not include information that can directly identify you without your specific consent.

Where can I get more information?

If you have any questions about the survey, please email us: aa1r21@soton.ac.uk
Abdullah Almontashiri.

What happens if there is a problem?

If you have a concern about any aspect of this study, you should speak to the researchers who will do their best to answer your questions.

If you remain unhappy or have a complaint about any aspect of this study, please contact the University of Southampton Research Integrity and Governance Manager (023 8059 5058, rgoinfo@soton.ac.uk).

Data Protection Privacy Notice

The University of Southampton conducts research to the highest standards of research integrity. As a publicly-funded organisation, the University has to ensure that it is in the public interest when we use personally-identifiable information about people who have agreed to take part in research. This means that when you agree to take part in a research study, we will use information about you in

the ways needed, and for the purposes specified, to conduct and complete the research project. Under data protection law, 'Personal data' means any information that relates to and is capable of identifying a living individual. The University's data protection policy governing the use of personal data by the University can be found on its website (<https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page>).

This Participant Information Sheet tells you what data will be collected for this project and whether this includes any personal data. Please ask the research team if you have any questions or are unclear what data is being collected about you.

Our privacy notice for research participants provides more information on how the University of Southampton collects and uses your personal data when you take part in one of our research projects and can be found at <http://www.southampton.ac.uk/assets/sharepoint/intranet/Is/Public/Research%20and%20Integrity%20Privacy%20Notice/Privacy%20Notice%20for%20Research%20Participants.pdf>

Any personal data we collect in this study will be used only for the purposes of carrying out our research and will be handled according to the University's policies in line with data protection law. If any personal data is used from which you can be identified directly, it will not be disclosed to anyone else without your consent unless the University of Southampton is required by law to disclose it.

Data protection law requires us to have a valid legal reason ('lawful basis') to process and use your Personal data. The lawful basis for processing personal information in this research study is for the performance of a task carried out in the public interest. Personal data collected for research will not be used for any other purpose.

For the purposes of data protection law, the University of Southampton is the 'Data Controller' for this study, which means that we are responsible for looking after your information and using it properly. The University of Southampton will keep identifiable information about you for 10 years after the study has finished after which time any link between you and your information will be removed.

For studies involving other recruitment sites the following information must be included:
the University of Southampton will keep identifiable information about you from this study [for 10 years after the study has finished/ until 2031]

To safeguard your rights, we will use the minimum personal data necessary to achieve our research study objectives. Your data protection rights – such as to access, change, or transfer such information - may be limited, however, in order for the research output to be reliable and accurate. The University will not do anything with your personal data that you would not reasonably expect.

If you have any questions about how your personal data is used, or wish to exercise any of your rights, please consult the University's data protection webpage (<https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page>) where you can make a request using our online form. If you need further assistance, please contact the University's Data Protection Officer (data.protection@soton.ac.uk).

Thank you.

Thank the individual for taking the time to read the information sheet and considering taking part in the research.

In this interview, we designed guideline questions as follow:

Questions	Justification
Part 1: Background and Experience	
Can you please tell me a little about yourself?	To collect demographic details
Why do you search for datasets in your field, and how do you relate dataset search to your area of research?	To ensure their understanding of the dataset search domain
Part 2: About your tasks	
Describe to me a recent project that required you to search for a dataset.	To understand the nature of the relationship between dataset search and their domains
What types of datasets do you use?	To understand the type of used datasets (e.g., structured, semi-structured, unstructured)
What types of data content do you work with?	To understand the nature of the data they use and work with
Part 3: Searching for datasets	
When you search for datasets, what methods do you use?	To understand the methods used for searching datasets, including techniques, websites, and tools
Why do you prefer using these specific methods?	To understand the reasons behind using these methods
How often do you search for datasets, and how much time do you typically expect to spend on it?	To understand the time spent and how current methods are used to retrieve datasets
Where do you usually search for datasets?	To explore the methods used for searching datasets (e.g., private repositories, public databases, journals, or personal contacts), which would help us understand the features of these methods
What information do you usually look for, and what considerations do you keep in mind when actively searching for data?	To identify the pieces of information that would assist them in selecting datasets
What challenges or issues do you face when searching for datasets?	To understand the obstacles in current methods and how they can be overcome
Part 4: Judging	
Can you describe how you judge each phase of searching for datasets?	To understand their methods for assessing the datasets they retrieved

Do you find metadata helpful when assessing datasets during your search? If yes, why. If no, why not.	To understand the usability of metadata in making judgments, which would provide insights into the methods used for making those judgments
Part 5: Quality	
What are some common data quality issues you encounter when searching for datasets?	To understand if there is a relationship between dataset search and data quality
What strategies or techniques do you use to ensure the quality of your datasets?	To understand how they currently address data quality when searching for datasets
Part 6: Important Information	
What are the most important elements or key information you focus on when searching for datasets?	To identify the elements they look for during a dataset search
What specific information do you need to know about a dataset before using it?	To understand what elements can facilitate dataset search
Part 7: Suggestions	
How would you imagine dataset search could be better?	To identify the aspects of dataset search that need improvement

TABLE A.1: Summary of Interview Questions and Justifications

Appendix B

Survey + Statistical Analysis

Participant Information Sheet

Study Title: Dataset search for biomedical researchers

Researcher: Abdullah Almontashiri
ERGO number: 92985

You are being invited to take part in the above research study. To help you decide whether you would like to take part or not, it is important that you understand why the research is being done and what it will involve. Please read the information below carefully and ask questions if anything is not clear or you would like more information before you decide to take part in this research. You may like to discuss it with others but it is up to you to decide whether or not to take part. If you are happy to participate you will be asked to sign a consent form.

What is the research about?

There are already several existing services and methods, such as Google Dataset Search that assist users in searching datasets; they include several new features to meet users' needs for datasets. Nevertheless, there is a lack of the methods of expressing requirements of datasets to be used in biomedical research domains. Due to the large number of datasets that exist on the web, researchers/scientists can encounter a number of difficulties when trying to discover appropriate datasets. The dataset search field is novel, and to date, there has been a lack of research on dataset search for biomedical research. The purpose of this research study is to supplement such efforts and develop new techniques that could help discover datasets for biomedical research. It will explore what types of additional information biomedical researchers need and how to identify datasets using provenance information in addition to their metadata.

Why have I been asked to participate?

The primary goal of this research is to confirm that extracting and providing provenance information of biomedical datasets in addition to their metadata can help to enhance the dataset search process. Therefore, you have been invited to participate since your knowledge of the field is sufficient and you are capable of answering the questions.

What will happen to me if I take part?

By responding to the survey, you will help us in the dataset search domain. Your answers will help to find out the requirements of dataset search for biomedical research. By responding to the survey, you need to read this form and then decide to agree or disagree. If your decision is to agree, you will start to answer some questions based on your experience in dataset search for biomedical purposes.

Are there any benefits in my taking part?

The participation in this study will be voluntary with entering a voucher raffle if they would like. The participants will contribute to develop the domain of dataset search for biomedical research. This contribution can improve the development of dataset search as well as facilitate the process of discovering datasets for biomedical purposes.

Are there any risks involved?

No. This study was approved by the Ethics committee at the University. There are no expected risks identified.

What data will be collected?

The aim of this study is to collect data to answer only research questions. Therefore, personal information will not be collected that could determine the identity of the participants. If you want to enter a raffle to win an Amazon voucher, we will ask you to provide your email address. Regarding demographic data, the area of research might be collected. Despite the possibility of collecting

personal information only for a raffle, the Data Protection Laws ensure the confidentiality of all collected data.

Will my participation be confidential?

Your participation and the information we collect about you during the course of the research will be kept strictly confidential.

Only members of the research team and responsible members of the University of Southampton may be given access to data about you for monitoring purposes and/or to carry out an audit of the study to ensure that the research is complying with applicable regulations. Individuals from regulatory authorities (people who check that we are carrying out the study correctly) may require access to your data. All of these people have a duty to keep your information, as a research participant, strictly confidential.

According to university policy, the collected data will be stored for ten years on the University server. All files will be encrypted to be more confidential. I intend to quote directly from the answers of the participant, if I need a direct quotation. Any audio records will be destroyed after the transcribing process is done. Then, the transcribed data will be stored with all the files of this study on the University server.

Do I have to take part?

No, it is entirely up to you to decide whether or not to take part. If you decide you want to take part, you will need to sign a consent form to show you have agreed to take part.

The participants, who will conduct the interview, will be asked to sign the consent form before starting the conversation.

What happens if I change my mind?

You have the right to change your mind and withdraw at any time without giving a reason and without your participant rights being affected.

All collected data in this survey will be discarded and deleted.

What will happen to the results of the research?

Your personal details will remain strictly confidential. Research findings made available in any reports or publications will not include information that can directly identify you without your specific consent.

Where can I get more information?

If you have any questions about the interviews, please email us: aa1r21@soton.ac.uk
Abdullah Almontashiri.

What happens if there is a problem?

If you have a concern about any aspect of this study, you should speak to the researchers who will do their best to answer your questions.

If you remain unhappy or have a complaint about any aspect of this study, please contact the University of Southampton Research Integrity and Governance Manager (023 8059 5058, rgoinfo@soton.ac.uk).

Data Protection Privacy Notice

The University of Southampton conducts research to the highest standards of research integrity. As a publicly-funded organisation, the University has to ensure that it is in the public interest when we use personally-identifiable information about people who have agreed to take part in research. This means that when you agree to take part in a research study, we will use information about you in the ways needed, and for the purposes specified, to conduct and complete the research project. Under data protection law, 'Personal data' means any information that relates to and is capable of identifying a living individual. The University's data protection policy governing the use of personal

data by the University can be found on its website
(<https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page>).

This Participant Information Sheet tells you what data will be collected for this project and whether this includes any personal data. Please ask the research team if you have any questions or are unclear what data is being collected about you.

Our privacy notice for research participants provides more information on how the University of Southampton collects and uses your personal data when you take part in one of our research projects and can be found at
<http://www.southampton.ac.uk/assets/sharepoint/intranet/Is/Public/Research%20and%20Integrity%20Privacy%20Notice/Privacy%20Notice%20for%20Research%20Participants.pdf>

Any personal data we collect in this study will be used only for the purposes of carrying out our research and will be handled according to the University's policies in line with data protection law. If any personal data is used from which you can be identified directly, it will not be disclosed to anyone else without your consent unless the University of Southampton is required by law to disclose it.

Data protection law requires us to have a valid legal reason ('lawful basis') to process and use your Personal data. The lawful basis for processing personal information in this research study is for the performance of a task carried out in the public interest. Personal data collected for research will not be used for any other purpose.

For the purposes of data protection law, the University of Southampton is the 'Data Controller' for this study, which means that we are responsible for looking after your information and using it properly. The University of Southampton will keep identifiable information about you for 10 years after the study has finished after which time any link between you and your information will be removed.

For studies involving other recruitment sites the following information must be included:
the University of Southampton will keep identifiable information about you from this study [for 10 years after the study has finished/ until 2031]

To safeguard your rights, we will use the minimum personal data necessary to achieve our research study objectives. Your data protection rights – such as to access, change, or transfer such information - may be limited, however, in order for the research output to be reliable and accurate. The University will not do anything with your personal data that you would not reasonably expect.

If you have any questions about how your personal data is used, or wish to exercise any of your rights, please consult the University's data protection webpage
(<https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page>) where you can make a request using our online form. If you need further assistance, please contact the University's Data Protection Officer (data.protection@soton.ac.uk).

Thank you.

Thank the individual for taking the time to read the information sheet and considering taking part in the research.



Consent

Study title: Dataset search for Biomedical Researchers

Ethics/ERGO no: 92985

This survey is a part of my research on dataset search for biomedical researchers. The questions are designed to explore how provenance information—that is, the origin and history of datasets—can improve the process of searching for biomedical datasets. Your responses will contribute valuable insights into whether provenance information effectively aids in selecting the biomedical datasets. This research is being conducted at the School of Electronics and Computer Science, University of Southampton, and is for only academic research purposes. The survey should take approximately 15-20 minutes to complete. Please be assured that your responses will be kept completely anonymous and confidential. Unless you would like to enter a raffle to win an Amazon voucher, we will ask you to enter your email. You have the right to withdraw from the survey at any point without penalty. Your time and your participation are appreciated.

If you have any questions about the survey, please email us: aa1r21@soton.ac.uk

Abdullah Almuntashiri.

Participant Information Sheet

Please tick this box to indicate that you consent to taking part in this survey.

- ☐ I agree
- ☐ I disagree

Demographic

What best describes you?

- ☐ Professor
- ☐ Lecturer
- ☐ Research Fellow
- ☐ Postdoctoral researcher
- ☐ PhD student
- ☐ Other:

What is your area of research?

Please briefly describe your role.

Block 2

Task 1: Imagine you are a new postgraduate biomedical researcher. You would like to initiate a project looking at factors contributing to Inflammatory Bowel Disease (IBD) in infants. You need a dataset for investigation purposes to understand this disease. As you search, you need to identify whether a particular dataset is worth downloading, extracting and evaluating for your use.

1: You submit a query on **datamed.org** to find a dataset to conduct the analysis and investigation.

Click **NEXT** to see the results

You see :

- **Name:** Inflammatory Bowel Disease Genetics
- **Repository:** NIDDK Central
- **Repository Identifier:** NIDDK.cr:IBD
- **Data or Study Types:** clinical trial
- **Source Organization:** National Institute of Diabetes and Digestive and Kidney Diseases
- **Access Conditions:** available with restriction
- **Access Hyperlink:** <https://www.niddkrepository.org/studies/icdb/IBD>.

Note: Downloading this dataset for in-depth inspection will take 90 minutes of your time.

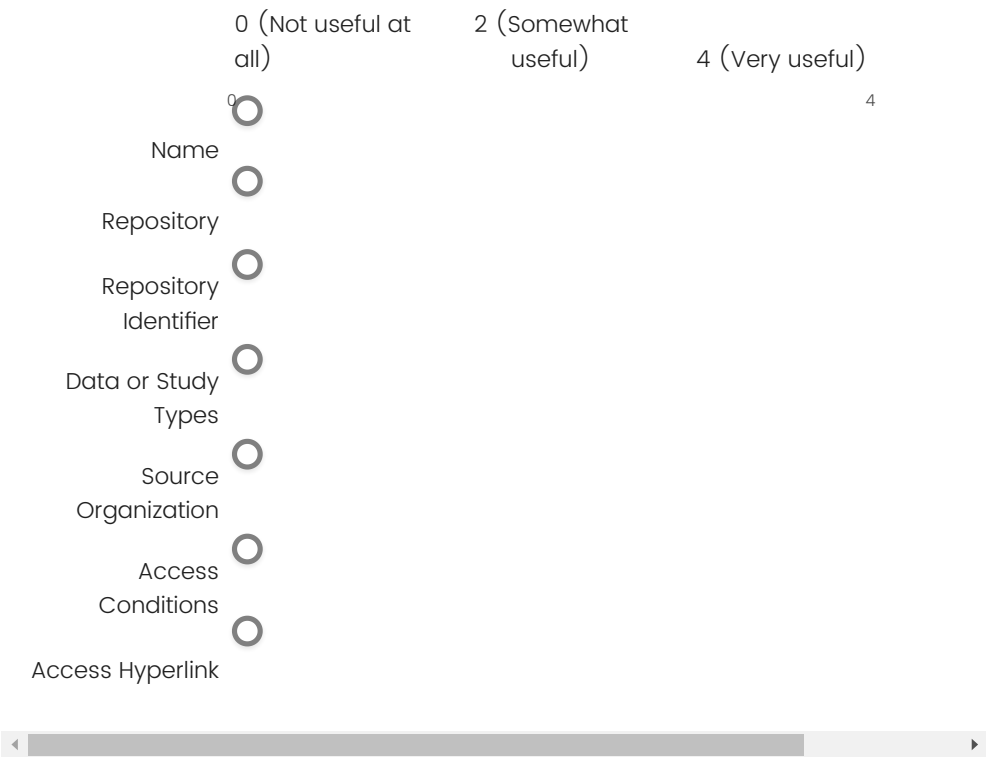


Q1: Would you download this dataset now or would you like to see more dataset options?

- ☐ No, I want to see more options
- ☐ Yes, download this one
- ☐ I'm not sure yet
- ☐ Other:

Q2: If yes, what part of the above information helped you decide?

Tip: Rank each element from 0 (Not useful at all) to 4 (Very useful)



Q3: If no, what are some reasons that you would like to see more dataset options?

Block 3

2: You submit a query on **pubmed.ncbi.nlm.nih.gov/** to expand you current dataset to complete your investigation

Click **NEXT** to see the results

Title: Inflammatory Bowel Disease in Children and Adolescents

Abstract: The inflammatory bowel diseases (IBDs), including ulcerative colitis and Crohn disease, are chronic inflammatory disorders of the gastrointestinal tract most often diagnosed in adolescence and young adulthood, with a rising incidence in pediatric populations. These disorders are common enough in children that most pediatricians and other pediatric clinicians will encounter children with IBD in their general practice. Inflammatory bowel disease is caused by a dysregulated mucosal immune response to the intestinal microflora in genetically predisposed hosts. Although children can present with the classic symptoms of weight loss, abdominal pain, and bloody diarrhea, many present with nonclassic symptoms of isolated poor growth, anemia, or other extraintestinal manifestations. Once IBD is diagnosed, the goals of therapy consist of eliminating symptoms, normalizing quality of life, restoring growth, and preventing complications while minimizing the adverse effects of medications. Unique considerations when treating children and adolescents with IBD include attention to the effects of the disease on growth and development, bone health, and psychosocial functioning. The purpose of this review is to provide a contemporary overview of the epidemiologic features, pathogenesis, diagnosis, and management of IBD in children and adolescents.

Note: Downloading this dataset for in-depth inspection will take 90 minutes of your time.



Q1: Would you download this dataset now or would you like to see more dataset options?

- ☐ No, I want to see more options
- ☐ Yes, download this one
- ☐ I'm not sure yet
- ☐ other:

Q2: If yes, what part of the above information helped you decide?

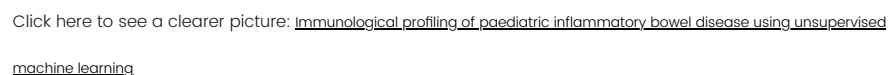
Tip: Rank each element from 0 (Not useful at all) to 4 (Very useful)

	0 (Not useful at all)	2 (Somewhat useful)	4 (Very useful)
Tilte	<input type="radio"/>		<input type="radio"/>
Abstract	<input type="radio"/>		<input type="radio"/>

--

3: You submit a query on **PubMed (new search portal)** for a new analysis.

Click **NEXT** to see the results



Note: Downloading this dataset for in-depth inspection will take 90 minutes of your time.

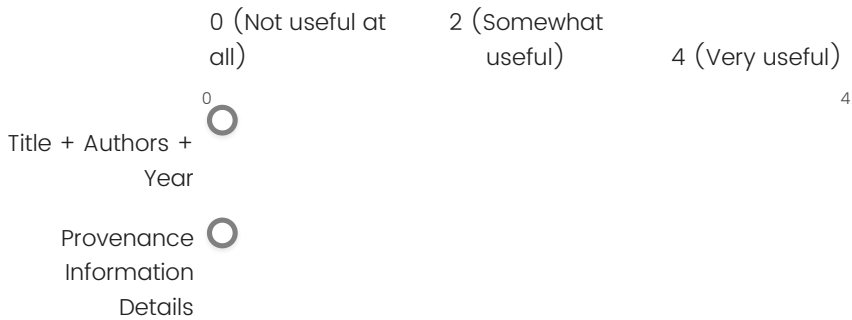


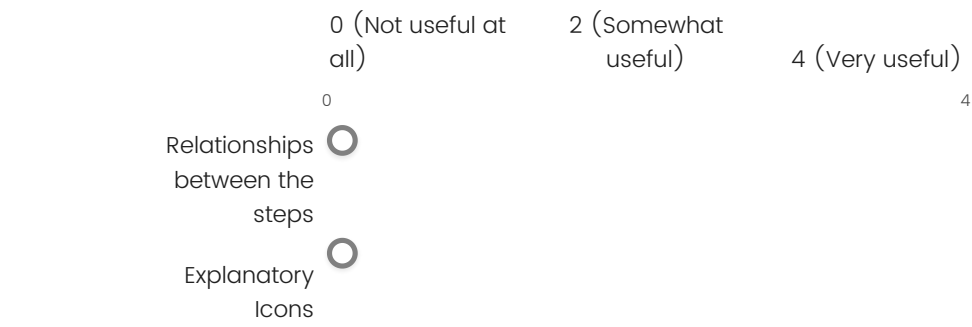
Q1: Would you download this dataset now or would you like to see more dataset options?

- ☐ No, I want to see more options
- ☐ Yes, download this one
- ☐ I'm not sure yet
- ☐ Other:

Q2: If yes, what part of the above information helped you decide?

Tip: Rank each element from 0 (Not useful at all) to 4 (Very useful)





Q2.B: Based on this information, check all the items you feel confident about.

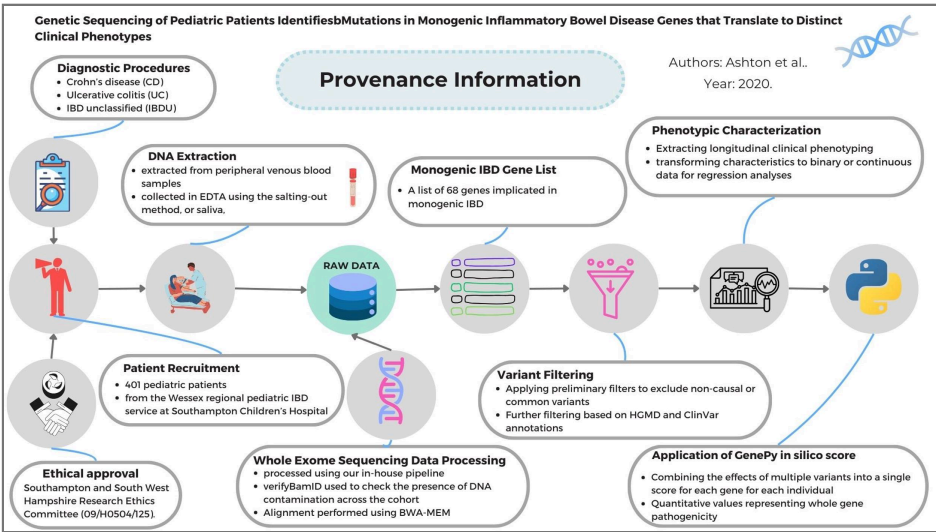
- ☐ Getting ethical approval
- ☐ How the data has been collected
- ☐ How the data has been processed
- ☐ Understanding the steps followed
- ☐ Understanding the outcome of this experiment
- ☐ Other:

Q3: If no, what are some reasons that you would like to see more dataset options?

Block 5

4: You submit a query on **PubMed (new search portal)** for a new analysis.

Click **NEXT** to see the results



Click here to see a clearer picture: [Genetic sequencing of pediatric patients](#)

Note: Downloading this dataset for in-depth inspection will take 90 minutes of your time.

Abstract:

Objectives: Monogenic inflammatory bowel disease (IBD) comprises rare Mendelian causes of gut inflammation, often presenting in infants with severe and atypical disease. This study aimed to identify clinically relevant variants within 68 monogenic IBD genes in an unselected pediatric IBD cohort.

Methods: Whole exome sequencing was performed on patients with pediatric-onset disease. Variants fulfilling the American College of Medical Genetics criteria as "pathogenic" or "likely pathogenic" were assessed against phenotype at diagnosis and follow-up. Individual patient variants were assessed and processed to generate a per-gene, per-individual, deleteriousness score.

Results: Four hundred one patients were included, and the median age of disease-onset was 11.92 years. In total, 11.5% of patients harbored a monogenic variant. TRIM22-related disease was implicated in 5 patients. A pathogenic mutation in the Wiskott-Aldrich syndrome (WAS) gene was confirmed in 2 male children with severe pancolonic inflammation and primary sclerosing cholangitis. In total, 7.3% of patients with Crohn's disease had apparent autosomal recessive, monogenic NOD2-related disease. Compared with non-NOD2 Crohn's disease, these patients had a marked stricturing phenotype (odds ratio 11.52, significant after correction for disease location) and had undergone significantly more intestinal resections (odds ratio 10.75). Variants in ADA, FERMT1, and LRBA did not meet the criteria for monogenic disease in any patients; however, case-control analysis of mutation burden significantly implicated these genes in disease etiology.

Discussion: Routine whole exome sequencing in pediatric patients with IBD results in a precise molecular diagnosis for a subset of patients with IBD, providing the opportunity to personalize therapy. NOD2 status informs risk of stricturing disease requiring surgery, allowing clinicians to direct prognosis and intervention.

Note: Downloading this dataset for in-depth inspection will take 90 minutes of your time.

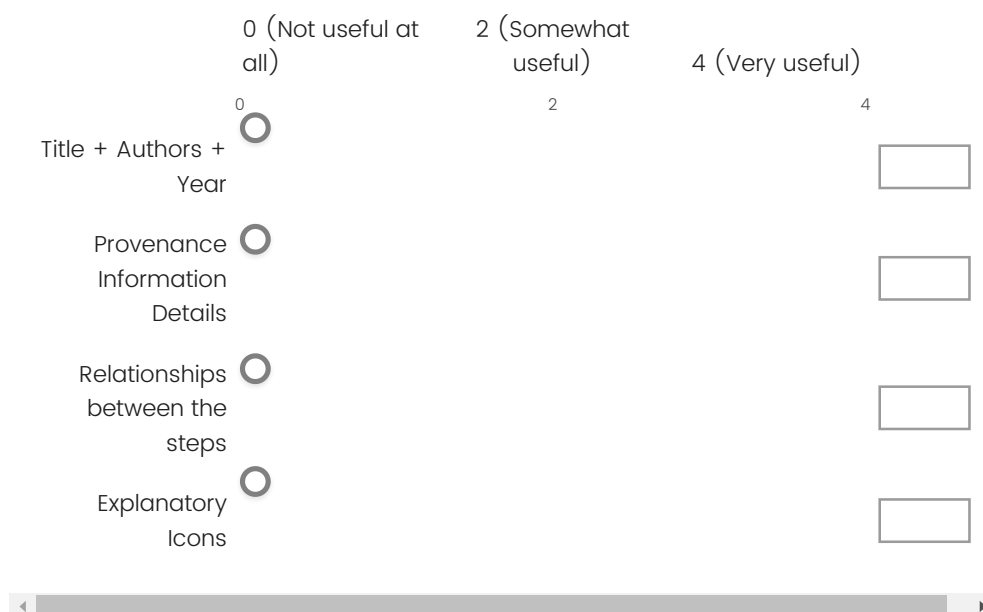


Q1: Would you download this dataset now or would you like to see more dataset options?

- ☐ No, I want to see more options
- ☐ Yes, download this one
- ☐ I'm not sure yet
- ☐ Other:

Q2: If yes, what part of the above information helped you decide?

Tip: Rank each element from 0 (Not useful at all) to 4 (Very useful)



Q2.B: Based on this information, check all the items you feel confident about.

- ☐ Getting ethical approval
- ☐ How the data has been collected
- ☐ How the data has been processed
- ☐ Understanding the steps followed
- ☐ Understanding the outcome of this experiment
- ☐ Other:

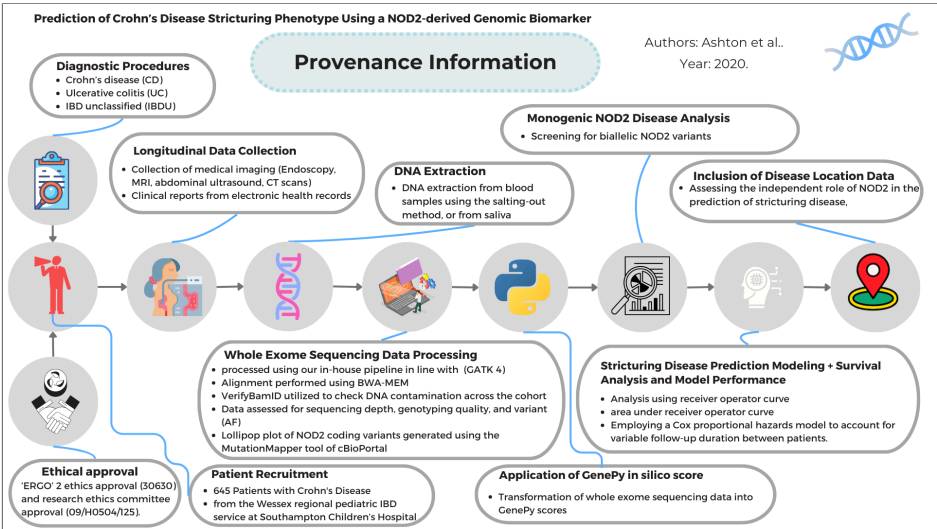
Q3: If no, what are some reasons that you would like to see more dataset options?

Block 6

Task 2: Imagine you are a research fellow in a biomedical research group. One part of this group is focused on Crohn's disease, including its history, causes and possible risks. You would like to initiate a project looking at factors contributing to Crohn's disease. Your task requires a dataset to investigate the impact factors of this disease. As you search for a dataset, you need to identify whether a particular dataset is worth downloading, extracting and analysing for fitness for your use.

1: You submit a query on **PubMed (new search portal)** to find a dataset to conduct the analysis and investigation.

Click **NEXT** to see the results



Click here to see a clearer picture: [Prediction of Crohn's Disease Structuring Phenotype Using a NOD2-derived Genomic Biomarker](#)

Note: Downloading this dataset for in-depth inspection will take 90 minutes of your time.

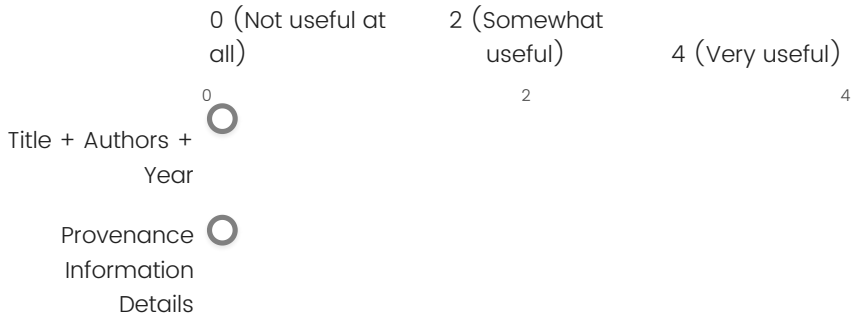


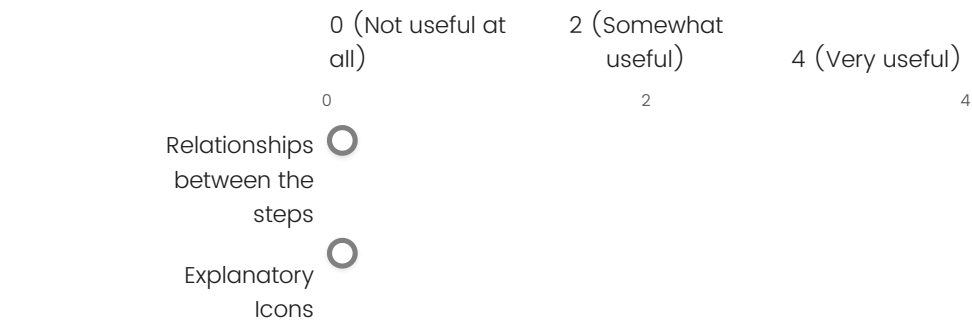
Q1: Would you download this dataset now or would you like to see more dataset options?

- ☐ No, I want to see more options
- ☐ Yes, download this one
- ☐ I'm not sure yet
- ☐ Other:

Q2: If yes, what part of the above information helped you decide?

Tip: Rank each element from 0 (Not useful at all) to 4 (Very useful)





Q2.B: Based on this information, check all the items you feel confident about.

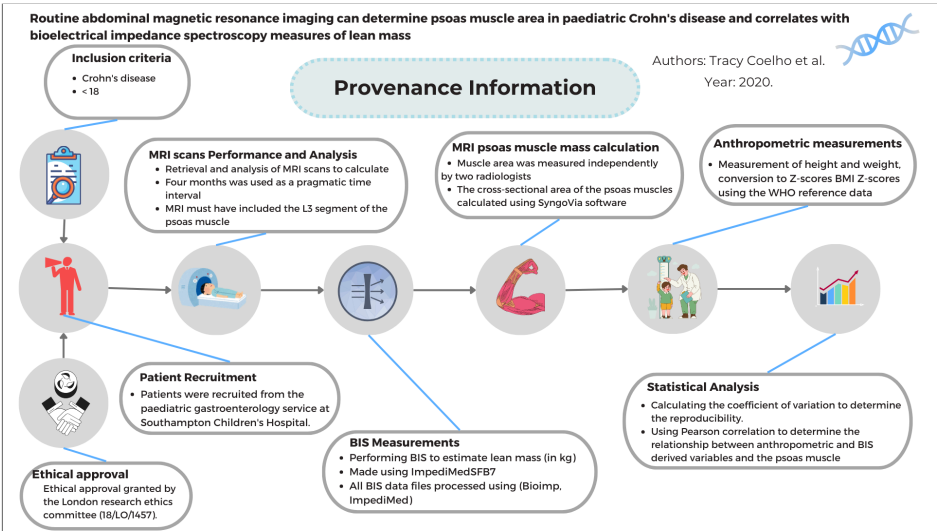
- ☐ Getting ethical approval
- ☐ How the data has been collected
- ☐ How the data has been processed
- ☐ Understanding the steps followed
- ☐ Understanding the outcome of this experiment
- ☐ Other:

Q3: What are some reasons that you would like to see more dataset options?

Block 7

2: You submit a query on **PubMed (new search portal)** to expand you current dataset to complete your investigation.

Click **NEXT** to see the results



Click here to see a clearer picture:[Routine abdominal magnetic resonance imaging can determine psoas muscle area in paediatric crohn's disease](#)

Abstract:

Background: Paediatric Crohn's disease (CD) has been associated with undernutrition. Accurate and accessible measures of body composition would provide data to personalise nutritional therapy. We assessed feasibility of MRI-derived measures of psoas cross-sectional area (PCSA) in paediatric CD and correlated with anthropometric and bioelectrical impedance spectroscopy (BIS) measures.

Methods: MRI small bowel/pelvis images of patients with CD, aged <18 years, were retrieved. Patients with concurrent anthropometric and BIS measurements were eligible for inclusion. The PCSA at L3 was calculated by two assessors and combined. To assess reproducibility of measures we calculated the coefficient of variation (CoV). Age, height-Z-scores, weight-Z-scores and BIS measures were correlated with PCSA. Using normal paediatric data from CT-scans we derived psoas area Z-scores for our cohort.

Results: 10 patients were included. Mean age at MRI scan was 14.6 years (11.7–16.3). PCSA was calculated for all MRI scans. There was high reproducibility between measurers, mean CoV 0.099. There was a significant positive correlation between PCSA and BIA-derived fat free mass, Pearson correlation coefficient (PCC) 0.831, $p = 0.003$. Correlation coefficients for PCSA and Height-for-age Z-

score, weight-for-age -Z-score and age were PCC 0.343- p = 0.33, PCC = 0.222- p = 0.54, and PCC 0.6034- p = 0.065, respectively. The mean PCSA Z-score was -1.81, with 70% of the patients having a Z-score < -2.0.

Conclusions: These data demonstrate the feasibility of deriving measures of body composition from routine MRI imagine. There was significant positive correlation between PCSA and BIS-derived lean mass. Further studies are required to confirm applicability of normal ranges prior to routine clinical implementation.

Note: Downloading this dataset for in-depth inspection will take 90 minutes of your time.

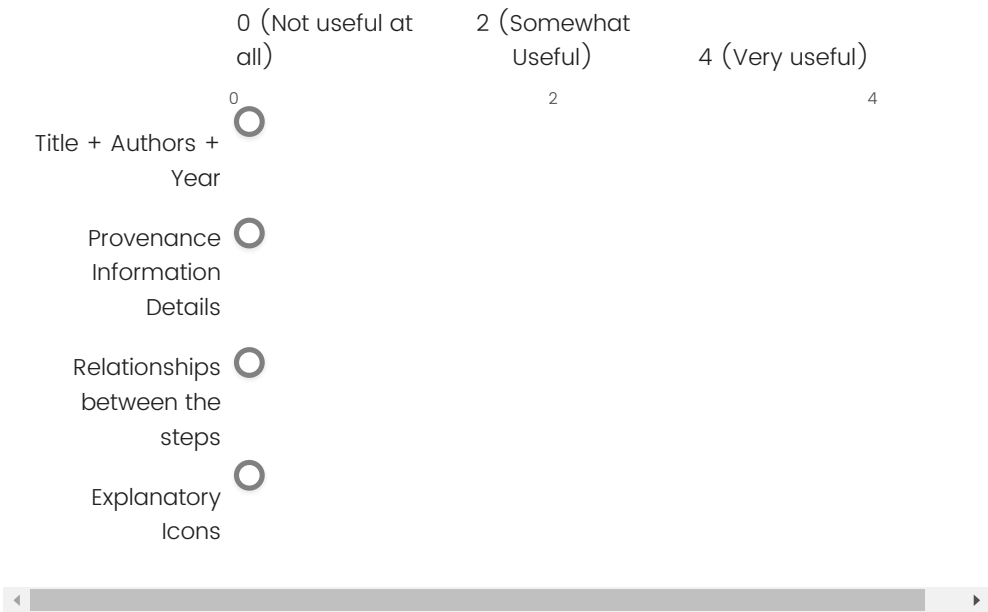


Q1: Would you download this dataset now or would you like to see more dataset options?

- ☐ No, I want to see more options
- ☐ Yes, download this one
- ☐ I'm not sure yet
- ☐ Other:

Q2: If yes, what part of the above information helped you decide?

Tip: Rank each element from 0 (Not useful at all) to 4 (Very useful)



Q2.B: Based on this information, check all the items you feel confident about.

- ☐ Getting ethical approval
- ☐ How the data has been collected
- ☐ How the data has been processed
- ☐ Understanding the steps followed
- ☐ Understanding the outcome of this experiment
- ☐ Other:

Q3: If no, what are some reasons that you would like to see more dataset options?

Block 8

3: You submit a query on **datamed.org** to integrate a dataset to yours for further evaluation.

Click **NEXT** to see the results

You see :

- **Name:** Ileal immune maturation in Pediatric Crohn's Disease
- **Repository:** Gene Expression Omnibus
- **Identifier:** geo.series:GSE62207
- **Description:** We report the global pattern of ileal gene expression in a cohort of 310 treatment-naïve pediatric Crohn Disease patients and controls. We focus on genes with consistent altered expression in the ileum of younger (Paris age A1a) vs older (Paris age A1b) patients.
- **Data or Study Types:** Expression profiling by high throughput sequencing
- **Source Organization:** National Center for Biotechnology Information
- **Access Conditions:** available
- **Year:** 2015
- **Access Hyperlink:** <http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GSE62207>

Note: Downloading this dataset for in-depth inspection will take 90 minutes of your time.



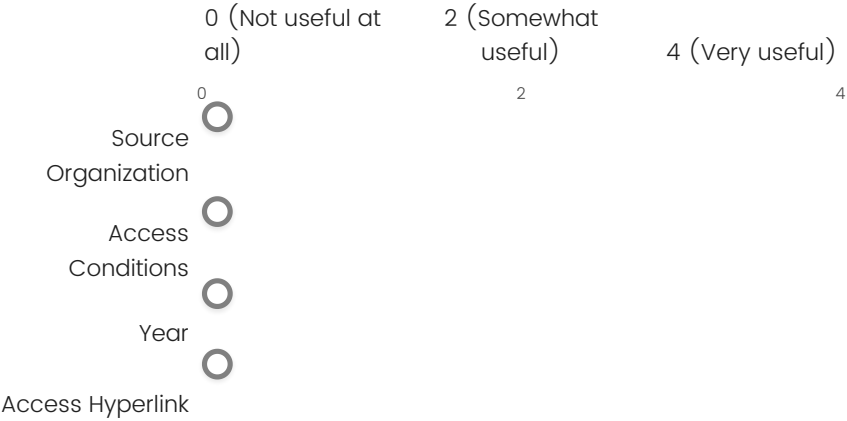
Q1: Would you download this dataset now or would you like to see more dataset options?

- ☐ No, I want to see more options
- ☐ Yes, download this one
- ☐ I'm not sure yet
- ☐ Other:

Q2: If yes, what part of the above information helped you decide?

Tip: Rank each element from 0 (Not useful at all) to 4 (Very useful)

	0 (Not useful at all)	2 (Somewhat useful)	4 (Very useful)
Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Repository	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Identifier	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Data or Study Types	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



Q3: If no, what are some reasons that you would like to see more dataset options?

Block 9

4: You submit a query on **pubmed.ncbi.nlm.nih.gov/** to expand your current dataset.

Click **NEXT** to see the results

Title: Elevated Levels of the Cytokine LIGHT in Pediatric Crohn's Disease

Abstract: LIGHT (homologous to lymphotoxins, exhibits inducible expression, and competes with HSV glycoprotein D for herpes virus entry mediator, a receptor expressed by T lymphocytes), encoded by the TNFSF14 gene, is a cytokine belonging to the TNF superfamily. On binding to its receptors, herpes virus entry mediator and lymphotoxin β receptor, it activates inflammatory responses. We conducted this study to determine whether plasma LIGHT levels are elevated in Crohn's disease (CD) in a pediatric population with the aim of nominating this cytokine as a therapeutic target. We used a single-

molecule immunoassay to determine the circulating levels of free LIGHT in plasma from pediatric patients with CD in our biobank (n = 183), a panel of healthy pediatric (n = 9) or adult (n = 22) reference samples, and pediatric biobank controls (n = 19). We performed correlational analyses between LIGHT levels and the clinical characteristics of the CD cohort, including age, Montreal classification, family history, medical/surgical therapy, and routine blood test parameters. LIGHT levels were greatly elevated in CD, with an average of 305 versus 32.4 pg/ml for controls from the biobank ($p < 0.0001$). The outside reference samples showed levels of 57 pg/ml in pediatric controls and 55 pg/ml in adults ($p < 0.0001$). We found a statistically significant correlation between white blood cell count and free LIGHT ($p < 0.046$). We conclude that free, soluble LIGHT is increased 5- to 10-fold in pediatric CD across an array of disease subtypes and characteristics.

Note: Downloading this dataset for in-depth inspection will take 90 minutes of your time.

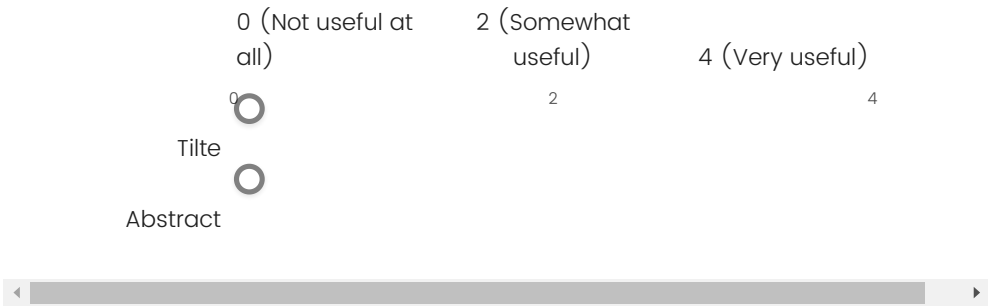


Q1: Would you download this dataset now or would you like to see more dataset options?

- ☐ No, I want to see more options
- ☐ Yes, download this one
- ☐ I'm not sure yet
- ☐ Other:

Q2: If yes, what part of the above information helped you decide?

Tip: Rank each element from 0 (Not useful at all) to 4 (Very useful)



Q3: If no, what are some reasons that you would like to see more dataset options?

Block 10

Thank you very much for taking time to complete this survey.
If you would like to discuss this project further please contact:
aarl21@soton.ac.uk

Write your email if you would like to enter a raffle to win an Amazon voucher.

Powered by Qualtrics

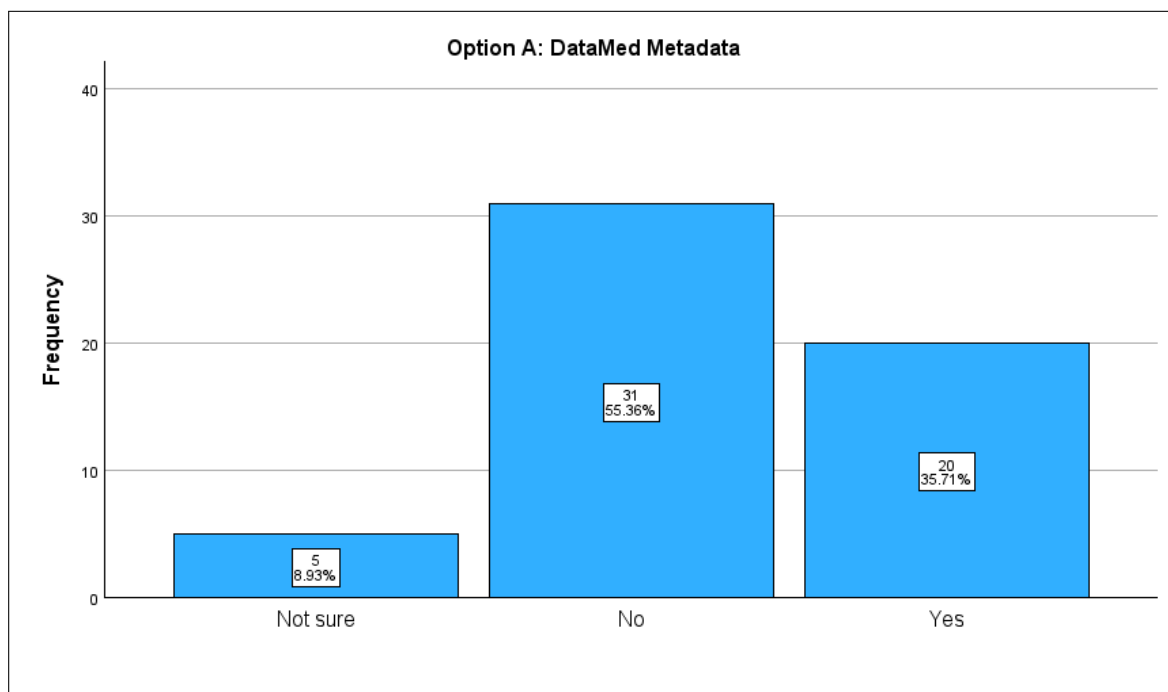


FIGURE B.1: Option A: DataMed Metadata

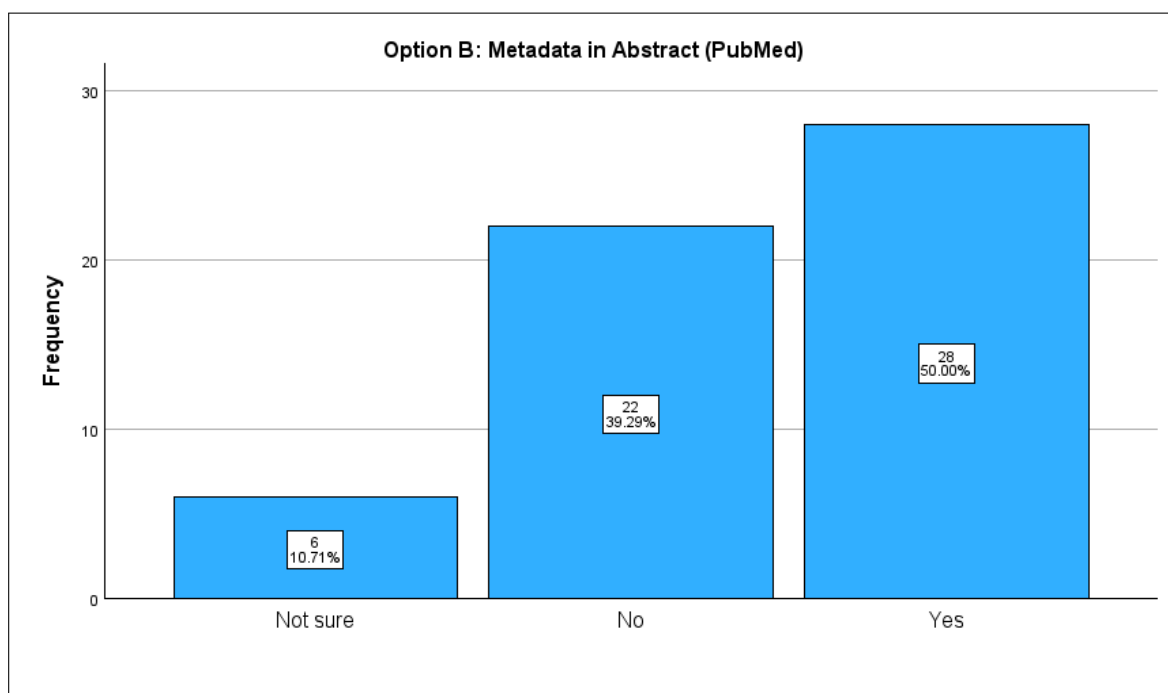


FIGURE B.2: Option B: Metadata in abstract

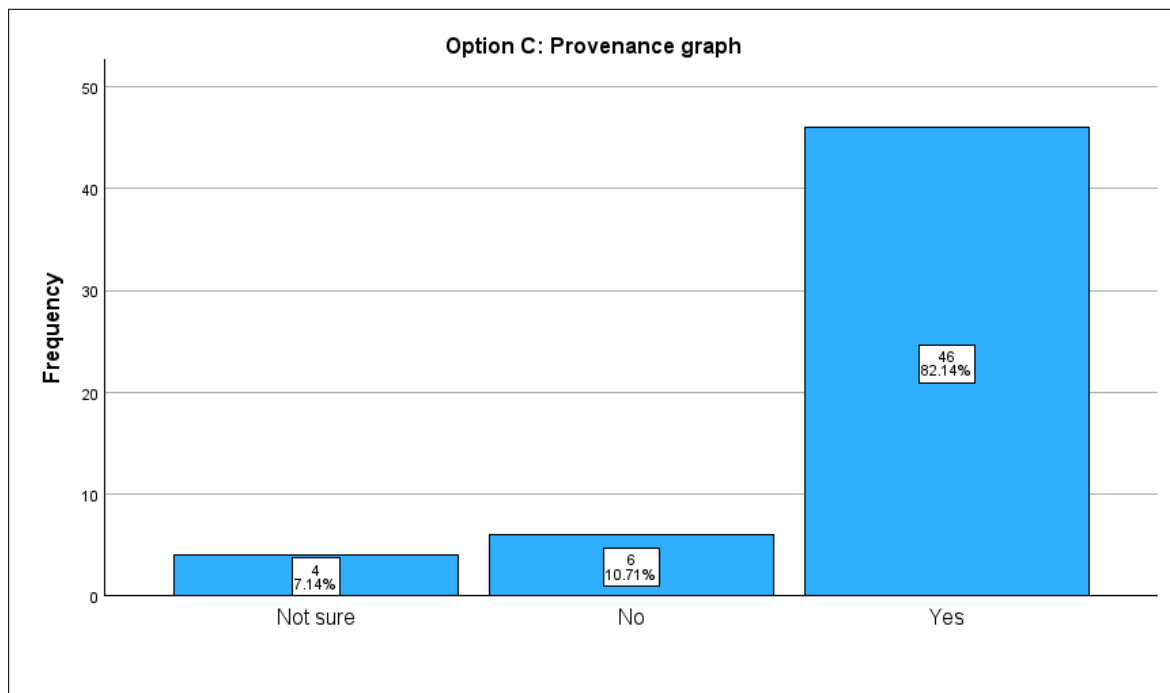


FIGURE B.3: Option C: Provenance metadata

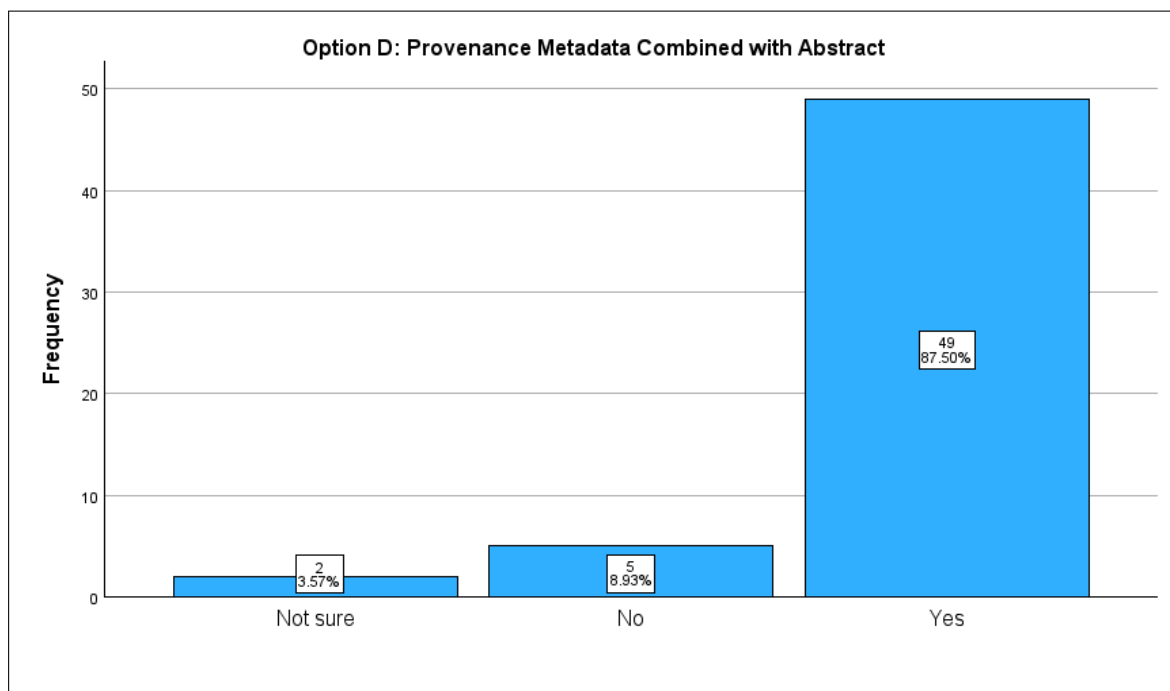


FIGURE B.4: Option D: Provenance metadata

Statistics						
		Ethics	Data_collected	Data_processed	Steps_followed	Outcome
N	Valid	46	46	46	46	46
	Missing	0	0	0	0	0
Mean		.67	.85	.85	.98	.89
Median		1.00	1.00	1.00	1.00	1.00
Mode		1	1	1	1	1
Range		1	1	1	1	1
Minimum		0	0	0	0	0
Maximum		1	1	1	1	1
Sum		31	39	39	45	41
Percentiles	25	.00	1.00	1.00	1.00	1.00
	50	1.00	1.00	1.00	1.00	1.00
	75	1.00	1.00	1.00	1.00	1.00

FIGURE B.5: Statistical tests conducted in SPSS 1

Statistics						
		Q78_6_TEXT	ResnifNO_B	SubjectiveAsses s1_C	SubjectiveAsses s2_C	EE
N	Valid	49	49	49	49	49
	Missing	0	0	0	0	0
Percentiles	25	.00	1.00	1.00	1.00	1.00
	50	1.00	1.00	1.00	1.00	1.00
	75	1.00	1.00	1.00	1.00	1.00

FIGURE B.6: Statistical tests conducted in SPSS 2

		Statistics				
		VAR00006	VAR00005	VAR00004	VAR00003	VAR00002
N	Valid	45	45	45	45	45
	Missing	0	0	0	0	0
Percentiles	25	.00	1.00	1.00	1.00	1.00
	50	1.00	1.00	1.00	1.00	1.00
	75	1.00	1.00	1.00	1.00	1.00

FIGURE B.7: Statistical tests conducted in SPSS 3

		Statistics				
		VAR00011	VAR00010	VAR00009	VAR00008	VAR00007
N	Valid	44	44	44	44	44
	Missing	0	0	0	0	0
Percentiles	25	.00	1.00	1.00	1.00	1.00
	50	1.00	1.00	1.00	1.00	1.00
	75	1.00	1.00	1.00	1.00	1.00

FIGURE B.8: Statistical tests conducted in SPSS 4

Chi-Square Tests

	Value	df	Asymptotic Significance (2- sided)
Pearson Chi-Square	9.827 ^a	12	.631
Likelihood Ratio	10.513	12	.571
N of Valid Cases	56		

a. 19 cells (90.5%) have expected count less than 5. The minimum expected count is .45.

FIGURE B.9: Statistical tests conducted in SPSS 5

Chi-Square Tests

	Value	df	Asymptotic Significance (2- sided)
Pearson Chi-Square	11.668 ^a	12	.473
Likelihood Ratio	12.389	12	.415
N of Valid Cases	56		

a. 19 cells (90.5%) have expected count less than 5. The minimum expected count is .54.

FIGURE B.10: Statistical tests conducted in SPSS 6

Chi-Square Tests

	Value	df	Asymptotic Significance (2- sided)
Pearson Chi-Square	15.536 ^a	12	.213
Likelihood Ratio	13.711	12	.320
N of Valid Cases	56		

a. 18 cells (85.7%) have expected count less than 5. The minimum expected count is .36.

FIGURE B.11: Statistical tests conducted in SPSS 7

Chi-Square Tests

	Value	df	Asymptotic Significance (2- sided)
Pearson Chi-Square	22.171 ^a	12	.036
Likelihood Ratio	17.335	12	.137
N of Valid Cases	56		

a. 17 cells (81.0%) have expected count less than 5. The minimum expected count is .18.

FIGURE B.12: Statistical tests conducted in SPSS 8

Independent Samples Test					
		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
T1O1_Num	Equal variances assumed	2.636	.107	-4.258	110
	Equal variances not assumed			-4.258	109.567

Independent Samples Test					
		t-test for Equality of Means			
		Significance		Mean Difference	Std. Error Difference
		One-Sided p	Two-Sided p		
T1O1_Num	Equal variances assumed	<.001	<.001	-.482	.113
	Equal variances not assumed	<.001	<.001	-.482	.113

FIGURE B.13: Statistical tests conducted in SPSS 9

Independent Samples Test					
		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
T1O2_Num	Equal variances assumed	27.632	<.001	-4.078	110
	Equal variances not assumed			-4.078	96.504

Independent Samples Test					
		t-test for Equality of Means			
		Significance		Mean Difference	Std. Error Difference
		One-Sided p	Two-Sided p		
T1O2_Num	Equal variances assumed	<.001	<.001	-.446	.109
	Equal variances not assumed	<.001	<.001	-.446	.109

FIGURE B.14: Statistical tests conducted in SPSS 10

		Value Label	N
Approach	1.00	T1O1_Num	56
	2.00	T1O2_Num	56
	3.00	T1O3_Num	56
	4.00	T1O4_Num	56

Dependent Variable: Response

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	13.299 ^a	3	4.433	13.234	<.001
Intercept	1476.004	1	1476.004	4406.197	<.001
Approaches	13.299	3	4.433	13.234	<.001
Error	73.696	220	.335		
Total	1563.000	224			
Corrected Total	86.996	223			

a. R Squared = .153 (Adjusted R Squared = .141)

FIGURE B.15: Statistical tests conducted in SPSS 11

<p>Dataset description in metadata – T1</p> <p>Reliability Statistics</p> <table> <tr> <th>Cronbach's Alpha</th><th>N of Items</th></tr> <tr> <td>.773</td><td>7</td></tr> </table>	Cronbach's Alpha	N of Items	.773	7	<p>Dataset description in metadata – T2</p> <p>Reliability Statistics</p> <table> <tr> <th>Cronbach's Alpha</th><th>N of Items</th></tr> <tr> <td>.892</td><td>8</td></tr> </table>	Cronbach's Alpha	N of Items	.892	8
Cronbach's Alpha	N of Items								
.773	7								
Cronbach's Alpha	N of Items								
.892	8								
<p>Dataset description in abstract – T1</p> <p>Reliability Statistics</p> <table> <tr> <th>Cronbach's Alpha</th><th>N of Items</th></tr> <tr> <td>.838</td><td>2</td></tr> </table>	Cronbach's Alpha	N of Items	.838	2	<p>Dataset description in abstract - T2</p> <p>Reliability Statistics</p> <table> <tr> <th>Cronbach's Alpha</th><th>N of Items</th></tr> <tr> <td>.760</td><td>2</td></tr> </table>	Cronbach's Alpha	N of Items	.760	2
Cronbach's Alpha	N of Items								
.838	2								
Cronbach's Alpha	N of Items								
.760	2								
<p>Provenance information – T1</p> <p>Reliability Statistics</p> <table> <tr> <th>Cronbach's Alpha</th><th>N of Items</th></tr> <tr> <td>.778</td><td>4</td></tr> </table>	Cronbach's Alpha	N of Items	.778	4	<p>Provenance information – T2</p> <p>Reliability Statistics</p> <table> <tr> <th>Cronbach's Alpha</th><th>N of Items</th></tr> <tr> <td>.817</td><td>4</td></tr> </table>	Cronbach's Alpha	N of Items	.817	4
Cronbach's Alpha	N of Items								
.778	4								
Cronbach's Alpha	N of Items								
.817	4								
<p>Provenance information with abstract – T1</p> <p>Reliability Statistics</p> <table> <tr> <th>Cronbach's Alpha</th><th>N of Items</th></tr> <tr> <td>.840</td><td>4</td></tr> </table>	Cronbach's Alpha	N of Items	.840	4	<p>Provenance information with abstract – T2</p> <p>Reliability Statistics</p> <table> <tr> <th>Cronbach's Alpha</th><th>N of Items</th></tr> <tr> <td>.818</td><td>4</td></tr> </table>	Cronbach's Alpha	N of Items	.818	4
Cronbach's Alpha	N of Items								
.840	4								
Cronbach's Alpha	N of Items								
.818	4								

FIGURE B.16: Statistical tests conducted in SPSS 12

	N	Mean	Std. Deviation	Std. Error Mean
DSName	20	3.05	1.146	.256
DSrepository	20	2.20	1.399	.313
RepositoryID	20	2.35	1.424	.319
DataType	20	3.00	1.076	.241
organization	20	3.10	.912	.204
DSAccess	20	2.40	1.392	.311
AccessLink	20	3.40	.821	.184

FIGURE B.17: Statistical tests conducted in SPSS 13

	N	Mean	Std. Deviation	Std. Error Mean
DSTtl	28	2.89	1.227	.232
DSAbst	28	2.96	1.290	.244

FIGURE B.18: Statistical tests conducted in SPSS 14

	N	Mean	Std. Deviation	Std. Error Mean
GnrInfo	46	2.78	1.365	.201
Prov	46	3.48	.888	.131
Rltn	46	3.17	1.338	.197
Icons	46	2.89	1.269	.187

FIGURE B.19: Statistical tests conducted in SPSS 15

	N	Mean	Std. Deviation	Std. Error Mean
GnrInfo	47	2.89	1.323	.193
Prov	48	3.46	.874	.126
Rltn	49	3.27	1.095	.156
Icons	48	3.02	1.296	.187

FIGURE B.20: Statistical tests conducted in SPSS 16

	N	Mean	Std. Deviation	Std. Error Mean
GnrInfo	45	2.67	1.552	.231
Prov	45	3.22	1.106	.165
Rltn	45	3.07	1.232	.184
Icons	45	2.91	1.328	.198

FIGURE B.21: Statistical tests conducted in SPSS 17

	N	Mean	Std. Deviation	Std. Error Mean
GnrInfo	43	2.67	1.426	.218
Prov	44	3.25	1.059	.160
Rltn	44	2.93	1.149	.173
Icons	44	2.93	1.301	.196

FIGURE B.22: Statistical tests conducted in SPSS 18

	N	Mean	Std. Deviation	Std. Error Mean
DsName	31	2.32	1.514	.272
DSRepos	31	2.84	1.319	.237
DSRepolD	31	3.06	1.263	.227
DSDataTyp	31	3.00	1.342	.241
DSOrgnzttn	31	2.52	1.546	.278
DSAccssC	31	2.97	1.494	.268
Yr	31	2.35	1.330	.239
Accsslink	31	2.87	1.565	.281

FIGURE B.23: Statistical tests conducted in SPSS 19

	N	Mean	Std. Deviation	Std. Error Mean
DSTtl	23	2.43	1.343	.280
DSAbst	24	3.17	1.049	.214

FIGURE B.24: Statistical tests conducted in SPSS 20

Appendix C

Human Study 3

Participant Information Sheet

Study Title: Improving biomedical dataset search with provenance information

Researcher: Abdullah Almuntashiri

ERGO number: 98745

You are being invited to take part in the above research study. To help you decide whether you would like to take part or not, it is important that you understand why the research is being done and what it will involve. Please read the information below carefully and ask questions if anything is not clear or you would like more information before you decide to take part in this research. You may like to discuss it with others but it is up to you to decide whether or not to take part. If you are happy to participate you will be asked to sign a consent form.

What is the research about?

There are already several existing services and methods, such as Google Dataset Search, that assist users in searching for datasets; they include several new features to meet users' needs. Nevertheless, there is a lack of methods for expressing the requirements of datasets to be used in biomedical research domains. Due to the large number of datasets available on the web, researchers and scientists can encounter numerous difficulties when trying to search appropriate datasets. The dataset search field is novel, and to date, there has been a lack of research on dataset search specifically for biomedical research. The purpose of this research study is to supplement existing efforts and develop new techniques to help select suitable datasets for biomedical research. It will assess the completeness of the provided provenance information and identify any missing elements. These new techniques will then be evaluated by biomedical researchers to determine their effectiveness.

Why have I been asked to participate?

The primary goal of this research is to assess whether the provided provenance information of biomedical datasets is complete. Therefore, you have been invited to participate since your knowledge of the field is sufficient and you are capable of answering the questions.

What will happen to me if I take part?

By conducting this interview, you will help us in the dataset search domain. Your answers will help to assess the completeness of the provided provenance information. Before conducting the interview, you need to read this form and then decide to agree or disagree. If your decision is to agree, you will start to answer some questions based on your experience in dataset search for biomedical purposes.

Are there any benefits in my taking part?

The participation in this study will be voluntary. The participants will contribute to develop the domain of dataset search for biomedical research. This contribution can improve the development of dataset search as well as facilitate the process of discovering datasets for biomedical purposes.

Are there any risks involved?

No. This study was approved by the Ethics committee at the University. There are no expected risks identified.

What data will be collected?

The aim of this study is to collect data to answer only research questions. Therefore, any personal information will be pseudonymised during transcription. Regarding demographic data, the area of research might be collected. Despite there is no personal information will be captured during transcription, all collected data will be anonymous and confidential with Data Protection Laws. I would like to record the interviews and then transcript them. Therefore, the participants will be informed that I intend to record the interview and this is optional; they have the right to agree or disagree. A tick box for the audio recording agreement will be provided in the consent form.

Will my participation be confidential?

[Date: 15-08-2024] [Version number: 2]

[Ethics/IRAS number: 98745]

Your participation and the information we collect about you during the course of the research will be kept strictly confidential.

Only members of the research team and responsible members of the University of Southampton may be given access to data about you for monitoring purposes and/or to carry out an audit of the study to ensure that the research is complying with applicable regulations. Individuals from regulatory authorities (people who check that we are carrying out the study correctly) may require access to your data. All of these people have a duty to keep your information, as a research participant, strictly confidential.

According to university policy, the collected data will be stored for ten years on the University server. All files will be encrypted to be more confidential. I intend to quote directly from the answers of the participant, if I need a direct quotation. Any audio records will be destroyed after the transcribing process is done. Then, the transcribed data will be stored with all the files of this study on the University server.

Do I have to take part?

No, it is entirely up to you to decide whether or not to take part. If you decide you want to take part, you will need to sign a consent form to show you have agreed to take part. The participants, who will conduct the interview, will be asked to sign the consent form before starting the conversation.

What happens if I change my mind?

You have the right to change your mind and withdraw at any time without giving a reason and without your participant rights being affected. All collected data in this survey will be discarded and deleted.

What will happen to the results of the research?

Your personal details will remain strictly confidential. Research findings made available in any reports or publications will not include information that can directly identify you without your specific consent.

Where can I get more information?

If you have any questions about the interviews, please email us: aa1r21@soton.ac.uk
Abdullah Almuntashiri.

What happens if there is a problem?

If you have a concern about any aspect of this study, you should speak to the researchers who will do their best to answer your questions.

If you remain unhappy or have a complaint about any aspect of this study, please contact the University of Southampton Research Integrity and Governance Manager (023 8059 5058, rgoinfo@soton.ac.uk).

Data Protection Privacy Notice

The University of Southampton conducts research to the highest standards of research integrity. As a publicly-funded organisation, the University has to ensure that it is in the public interest when we use personally-identifiable information about people who have agreed to take part in research. This means that when you agree to take part in a research study, we will use information about you in the ways needed, and for the purposes specified, to conduct and complete the research project. Under data protection law, 'Personal data' means any information that relates to and is capable of identifying a living individual. The University's data protection policy governing the use of personal data by the University can be found on its website (<https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page>).

This Participant Information Sheet tells you what data will be collected for this project and whether this includes any personal data. Please ask the research team if you have any questions or are unclear what data is being collected about you.

Our privacy notice for research participants provides more information on how the University of Southampton collects and uses your personal data when you take part in one of our research projects and can be found at

<http://www.southampton.ac.uk/assets/sharepoint/intranet/Is/Public/Research%20and%20Integrity%20Privacy%20Notice/Privacy%20Notice%20for%20Research%20Participants.pdf>

Any personal data we collect in this study will be used only for the purposes of carrying out our research and will be handled according to the University's policies in line with data protection law. If any personal data is used from which you can be identified directly, it will not be disclosed to anyone else without your consent unless the University of Southampton is required by law to disclose it.

Data protection law requires us to have a valid legal reason ('lawful basis') to process and use your Personal data. The lawful basis for processing personal information in this research study is for the performance of a task carried out in the public interest. Personal data collected for research will not be used for any other purpose.

For the purposes of data protection law, the University of Southampton is the 'Data Controller' for this study, which means that we are responsible for looking after your information and using it properly. The University of Southampton will keep identifiable information about you for 10 years after the study has finished after which time any link between you and your information will be removed.

For studies involving other recruitment sites the following information must be included:

the University of Southampton will keep identifiable information about you from this study [for 10 years after the study has finished/ until 2031]

To safeguard your rights, we will use the minimum personal data necessary to achieve our research study objectives. Your data protection rights – such as to access, change, or transfer such information - may be limited, however, in order for the research output to be reliable and accurate. The University will not do anything with your personal data that you would not reasonably expect.

If you have any questions about how your personal data is used, or wish to exercise any of your rights, please consult the University's data protection webpage (<https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page>) where you can make a request using our online form. If you need further assistance, please contact the University's Data Protection Officer (data.protection@soton.ac.uk).

Thank you for taking the time to read the information sheet and considering taking part in the research.

References

- Samuel Abedu, Ahmad Abdellatif, and Emad Shihab. Llm-based chatbots for mining software repositories: Challenges and opportunities. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, pages 201–210, 2024.
- Mansoor Ahmed, Amil Rohani Dar, Markus Helfert, Abid Khan, and Jungsuk Kim. Data provenance in healthcare: Approaches, challenges, and future directions. *Sensors*, 23(14): 6495, 2023.
- Uchenna Akujuobi and Xiangliang Zhang. Delve: a dataset-driven scholarly search and analysis system. *ACM SIGKDD Explorations Newsletter*, 19(2):36–46, 2017.
- Amro Al-Said Ahmad and Peter Andras. Scalability analysis comparisons of cloud-based software services. *Journal of Cloud Computing*, 8(1):10, 2019.
- Abdullah Hamed Almuntashiri, Luis-Daniel Ibáñez, and Adriane Chapman. A taxonomy of dataset search. In *The International Conference of Advanced Computing and Informatics*, pages 562–573. Springer, 2022.
- Abdullah Hamed Almuntashiri, Luis-Daniel Ibáñez, and Adriane Chapman. Llms for the post-hoc creation of provenance. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 562–566. IEEE, 2024.
- Abdullah Hamed Almuntashiri, Luis-Daniel Ibáñez, and Adriane Chapman. Using llms to infer provenance information. In *Proceedings of ProvenanceWeek (PW’25)*, Berlin, Germany, June 22–27 2025. ACM. . URL <https://doi.org/10.1145/3736229.3736261>.
- Alawi A Alsheikh-Ali, Waqas Qureshi, Mouaz H Al-Mallah, and John PA Ioannidis. Public availability of published research data in high-impact journals. *PloS one*, 6(9):e24357, 2011.
- Micah Altman, Eleni Castro, Mercè Crosas, Philip Durbin, Alex Garnett, and Jen Whitney. Open journal systems and dataverse integration–helping journals to upgrade data publication for reusable research. *Code4Lib Journal*, 1(30), 2015.

- Nicholas R Anderson, E Sally Lee, J Scott Brockenbrough, Mark E Minie, Sherrilynne Fuller, James Brinkley, and Peter Tarczy-Hornoch. Issues in biomedical research data management and analysis: needs and barriers. *Journal of the American Medical Informatics Association*, 14(4):478–488, 2007.
- James Ashton, Enrico Mossotto, Imogen Stafford, Rachel Haggarty, Tracy Coelho, Akshay Batra, Nadeem A. Afzal, Matthew Mort, David Bunyan, R. Mark Beattie, and Sarah Ennis. Genetic sequencing of pediatric patients identifies mutations in monogenic inflammatory bowel disease genes that translate to distinct clinical phenotypes. *Journal of Pediatric Gastroenterology and Nutrition*, 70(3):e67–e76, 2020.
- James Ashton, Guo Cheng, Imogen S. Stafford, M. Kellermann, Eleanor Seaby, J.R. Fraser Cummings, Tracy Coelho, Akshay Batra, Nadeem A. Afzal, R. Mark Beattie, and Sarah Ennis. Prediction of crohn’s disease stricturing phenotype using a nod2-derived genomic biomarker. *Clinical Gastroenterology and Hepatology*, 21(4):890–898, 2023.
- Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 554–565. IEEE, 2019.
- Donald Ray Back. *Neoliberalism, Academic Capitalism and Higher Education in Developing Countries: The Case of Iraqi Kurdistan*. PhD thesis, Virginia Polytechnic Institute and State University, 2016.
- Tanya Barrett, Karen Clark, Robert Gevorgyan, Vyacheslav Gorelenkov, Eugene Gribov, Ilene Karsch-Mizrachi, Michael Kimelman, Kim D Pruitt, Sergei Resenchuk, Tatiana Tatusova, et al. Bioproject and biosample databases at ncbi: facilitating capture and organization of metadata. *Nucleic acids research*, 40(D1):D57–D63, 2012.
- Benjamin Baum, Christian R Bauer, Thomas Franke, Harald Kusch, Marcel Parciak, Thorsten Rottmann, Nadine Umbach, and Ulrich Sax. Opinion paper: Data provenance challenges in biomedical research. *IT-Information Technology*, 59(4):191–196, 2017.
- Jannis Beese, M Kazem Haki, Stephan Aier, and Robert Winter. Simulation-based research in information systems: epistemic implications and a review of the status quo. *Business & Information Systems Engineering*, 61:503–521, 2019.
- Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
- Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 41(D1):D36–D42, 2012.
- Ronald A Berk. Importance of expert judgment in content-related validity evidence. *Western journal of nursing research*, 12(5):659–671, 1990.
- Shitij Bhargava, Tsung-Ting Kuo, Ankit Goyal, Vincent Kuri, Gordon Lin, and Chun-Nan Hsu. biopdfx: preparing pdf scientific articles for biomedical text mining. Technical report, PeerJ Preprints, 2017.
- UK Biobank. About uk biobank, 2014.
- Dorothy Bishop, Doreen Cantrell, Peter Johnson, Shitij Kapur, Malcom Macleod, Caroline Savage, Jim Smith, S Tavaré, M Welham, J Williams, et al. Reproducibility and reliability of biomedical research: improving research practise. In *The Academy of Medical Sciences, Symposium report*, 2015.
- Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia-a crystallization point for the web of data. *Journal of web semantics*, 7(3):154–165, 2009.
- Andrew Black and Peter van Nederpelt. Dimensions of data quality (ddq): research paper. *DAMA NL Foundation*, pages 1–113, 2020.
- Tom Blount, Adriane Chapman, Michael Johnson, and Bertram Ludascher. Observed vs. possible provenance (research track). In *13th International Workshop on Theory and Practice of Provenance (TaPP 2021)*, 2021.
- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- Christine L. Borgman. *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press, Cambridge, MA, 2015. ISBN 9780262028561.
- Alexei Botchkarev. A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, 14:045–076, 2019.

- Mohamed Reda Bouadjene and Karin Verspoor. Multi-field query expansion is effective for biomedical dataset retrieval. *Database*, 2017:bax062, 2017.
- Keith M Bower. When to use fisher's exact test. In *American Society for Quality, Six Sigma Forum Magazine*, volume 2, pages 35–37. American Society for Quality Milwaukee, WI, USA, 2003.
- Petra M Boynton and Trisha Greenhalgh. Selecting, designing, and developing your questionnaire. *Bmj*, 328(7451):1312–1315, 2004.
- Ian Brace. *Questionnaire design: How to plan, structure and write survey material for effective market research*. Kogan Page Publishers, 2018.
- Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.
- Virginia Braun and Victoria Clarke. *Thematic analysis*. American Psychological Association, 2012.
- Virginia Braun and Victoria Clarke. To saturate or not to saturate? questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative research in sport, exercise and health*, 13(2):201–216, 2021.
- Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A Ball, Helen C Causton, et al. Minimum information about a microarray experiment (miame)—toward standards for microarray data. *Nature genetics*, 29(4):365–371, 2001.
- Kate Bredbenner and Martin Simon. The rise of the graphical abstract. *eLife*, 11:e77662, 2022.
- Dan Brickley, Matthew Burgess, and Natasha Noy. Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference*, pages 1365–1375, 2019.
- Giovanni Briganti. How chatgpt works: a mini review. *European Archives of Oto-Rhino-Laryngology*, 281(3):1565–1569, 2024.
- Svend Brinkmann and Steinar Kvale. *Doing interviews*, volume 2. Sage, 2018.
- Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM New York, NY, USA, 2002.

- T Brown, B Mann, N Ryder, M Subbiah, JD Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, et al. Language models are few-shot learners advances in neural information processing systems 33. 2020.
- Peter Buneman and Wang-Chiew Tan. Provenance in databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 1171–1173, 2007.
- Paul Cairns and Anna L Cox. *Research methods for human-computer interaction*, volume 10. Cambridge University Press Cambridge, 2008.
- Kathi Canese and Sarah Weis. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1), 2013.
- Mengyuan Cao, Hang Wang, Xiaoming Liu, Jiahao Wu, and Mengting Zhao. Llm collaboration plm improves critical information extraction tasks in medical articles. In *China Health Information Processing Conference*, pages 178–185. Springer, 2023.
- Christopher J Cardinale, Debra J Abrams, Frank D Mentch, John A Cardinale, Xiang Wang, Charlly Kao, Patrick Sleiman, and Hakon Hakonarson. Elevated levels of the cytokine light in pediatric crohn’s disease. *The Journal of Immunology*, 210(5):590–594, 2023.
- David Carmel, Naama Zwerdling, Ido Guy, Shila Ofek-Koifman, Nadav Har’el, Inbal Ronen, Erel Uziel, Sivan Yogev, and Sergey Chernov. Personalized social search based on the user’s social network. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM ’09*, page 1227–1236, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585123. . URL <https://doi.org/10.1145/1645953.1646109>.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.
- Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. Dataset search: a survey. *The VLDB Journal*, 29(1): 251–272, 2020.
- Adriane Chapman, Luca Lauro, Paolo Missier, and Riccardo Torlone. Dpds: assisting data science with data provenance. *Proceedings of the VLDB Endowment*, 15(12):3614–3617, 2022.
- You-Wei Cheah and Beth Plale. Provenance quality assessment methodology and framework. *Journal of Data and Information Quality (JDIQ)*, 5(3):1–20, 2014.

- Jinchi Chen, Xiaxia Wang, Gong Cheng, Evgeny Kharlamov, and Yuzhong Qu. Towards more usable dataset search: From query characterization to snippet generation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2445–2448, 2019.
- Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Hongming Chen, and Zhangmin Niu. An extensive benchmark study on biomedical text generation and mining with chatgpt. *Bioinformatics*, 39(9):btad557, 2023.
- Xiang Chen, Chaoyang Gao, Chunyang Chen, Guangbei Zhang, and Yong Liu. An empirical study on challenges for llm developers. *arXiv preprint arXiv:2408.05002*, 2024.
- Xiaoling Chen, Anupama E Gururaj, Burak Ozyurt, Ruiling Liu, Ergin Soysal, Trevor Cohen, Firat Tiryaki, Yueling Li, Nansu Zong, Min Jiang, et al. Datamed—an open source discovery index for finding biomedical datasets. *Journal of the American Medical Informatics Association*, 25(3):300–308, 2018.
- Zhiyu Chen. Dataset search and augmentation, 2022.
- James Cheney, Laura Chiticariu, Wang-Chiew Tan, et al. Provenance in databases: Why, how, and where. *Foundations and Trends® in Databases*, 1(4):379–474, 2009.
- Zijun Cheng, Qiujuan Lv, Jinyuan Liang, Yan Wang, Degang Sun, Thomas Pasquier, and Xueyuan Han. Kairos: Practical intrusion detection and investigation using whole-system provenance. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 3533–3551. IEEE, 2024.
- Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. Transformers: “the end of history” for natural language processing? In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 677–693. Springer, 2021.
- Fernando Chirigati, Rémi Rampin, Dennis Shasha, and Juliana Freire. Reprozip: Computational reproducibility with ease. In *Proceedings of the 2016 international conference on management of data*, pages 2085–2088, 2016.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.
- Emily Clough and Tanya Barrett. The gene expression omnibus database. In *Statistical genomics*, pages 93–110. Springer, 2016.

- William G. Cochran. *Sampling Techniques*. Wiley, New York, NY, 1963.
- Tracy Coelho, Enrico Mossotto, Yifang Gao, Rachel Haggarty, James J. Ashton, Akshay Batra, Imogen S. Stafford, Robert M. Beattie, Anthony P. Williams, and Sarah Ennis. Immunological profiling of paediatric inflammatory bowel disease using unsupervised machine learning. *Frontiers in Pediatrics*, 8:1–12, 2020.
- Aaron M Cohen and William R Hersh. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71, 2005.
- Sarah Cohen, Werner Nutt, and Yehoshua Sagie. Deciding equivalences among conjunctive aggregate queries. *Journal of the ACM*, 54(2), April 2007. . URL <http://doi.acm.org/10.1145/1219092.1219093>.
- Trevor Cohen, Kirk Roberts, Anupama E Gururaj, Xiaoling Chen, Saeid Pournajati, George Alter, William R Hersh, Dina Demner-Fushman, Lucila Ohno-Machado, and Hua Xu. A publicly available benchmark for biomedical dataset retrieval: the reference standard for the 2016 biocaddie dataset retrieval challenge. *Database*, 2017:bax061, 2017.
- Francis S Collins and Lawrence A Tabak. Policy: Nih plans to enhance reproducibility. *Nature*, 505(7485):612–613, 2014.
- Rory Collins. What makes uk biobank special? *Lancet (London, England)*, 379(9822):1173–1174, 2012.
- Donald C Comeau, Chih-Hsuan Wei, Rezarta Islamaj Doğan, and Zhiyong Lu. Pmc text mining subset in bioc: about three million full-text articles and growing. *Bioinformatics*, 35(18): 3533–3535, 2019.
- Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022.
- John W Creswell and Vicki L Plano Clark. *Designing and conducting mixed methods research*. Sage publications, 2017.
- John W Creswell and J David Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.
- V Curcin. Embedding data provenance into the learning health system to facilitate reproducible research (vol. 1 (2), p. e10019), 2017.
- Jan Czarnecki, Irene Nobeli, Adrian M Smith, and Adrian J Shepherd. A text-mining system for extracting metabolic reactions from full-text articles. *BMC bioinformatics*, 13(1):172, 2012.

- Sérgio Manuel Serra da Cruz, Marcos Bacis Ceddia, Renan Carvalho Távora Miranda, Gabriel Rizzo, Filipe Klinger, Renato Cerceau, Ricardo Mesquita, Ricardo Cerceau, Elton Carneiro Marinho, Eber Assis Schmitz, et al. Data provenance in agriculture. In *Provenance and Annotation of Data and Processes: 7th International Provenance and Annotation Workshop, IPAW 2018, London, UK, July 9-10, 2018, Proceedings*, pages 257–261. Springer, 2018.
- John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024.
- Tom De Nies, Sam Coppens, Erik Mannens, and Rik Van de Walle. Modeling uncertain provenance and provenance of uncertainty in w3c prov. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 167–168, 2013a.
- Tom De Nies, Sara Magliacane, Ruben Verborgh, Sam Coppens, Paul Groth, Erik Mannens, and Rik Van de Walle. Git2prov: Exposing version control system content as w3c prov. In *ISWC (Posters & Demos)*, pages 125–128, 2013b.
- Gero Decker, Alexander Grosskopf, and Alistair Barros. A graphical notation for modeling complex events in business processes. In *11th IEEE international enterprise distributed object computing conference (EDOC 2007)*, pages 27–27. IEEE, 2007.
- Jamie DeCoster and Heather Claypool. Data analysis in spss. 2004.
- John E Dennis Jr, David M Gay, and Roy E Walsh. An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software (TOMS)*, 7(3):348–368, 1981.
- Daniel Deutch, Yuval Moskovitch, and Val Tannen. A provenance framework for data-dependent process analysis. *Proceedings of the VLDB Endowment*, 7(6):457–468, 2014.
- Daniel Deutch, Amir Gilad, and Yuval Moskovitch. selp: selective tracking and presentation of data provenance. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1484–1487. IEEE, 2015.
- Kerry Dhakal. Nvivo. *Journal of the Medical Library Association: JMLA*, 110(2):270, 2022.
- Paolo et al. Di Tommaso. Nextflow enables reproducible computational workflows. *Future Generation Computer Systems*, 75:284–298, 2017. .
- Chaohai Ding, Mike Wald, and Gary Wills. A survey of open accessibility data. In *proceedings of the 11th web for all conference*, pages 1–4, 2014.

- Ram Dixit, Deevakar Rogith, Vidya Narayana, Mandana Salimi, Anupama Gururaj, Lucila Ohno-Machado, Hua Xu, and Todd R Johnson. User needs analysis and usability assessment of datamed—a biomedical data discovery index. *Journal of the American Medical Informatics Association*, 25(3):337–344, 2018.
- Xiao Dong, Yaoyun Zhang, and Hua Xu. Search datasets in literature: a case study of gwas. *AMIA Summits on Translational Science Proceedings*, 2017:40, 2017.
- Zoltán Dörnyei and Tatsuya Taguchi. *Questionnaires in second language research: Construction, administration, and processing*. Routledge, 2009.
- John Dudovskiy. The ultimate guide to writing a dissertation in business studies: A step-by-step assistance. *Pittsburgh, USA*, 51, 2016.
- Ruth Duerr. Data archives and repositories. In *Encyclopedia of Remote Sensing*, pages 127–131. Springer, 2014.
- Erik Duval, Wayne Hodgins, Stuart Sutton, and Stuart L Weibel. Metadata principles and practicalities. *D-lib Magazine*, 8(4):1–10, 2002.
- Ramtin Ehsani, Sakshi Pathak, and Preetha Chatterjee. Towards detecting prompt knowledge gaps for improved llm-guided issue resolution. *arXiv preprint arXiv:2501.11709*, 2025.
- Sabit Ekin. Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices. *Authorea Preprints*, 2023.
- Jian-Bing Fan, John Quackenbush, and Bart Wacek. Accelerating genomic data publishing and sharing. *Genomics Data*, 1:1, 2013.
- Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.
- Lisa M Federer, Ya-Ling Lu, Douglas J Joubert, Judith Welsh, and Barbara Brandys. Biomedical data sharing and reuse: attitudes and practices of clinical and scientific research staff. *PloS one*, 10(6):e0129506, 2015.
- William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.
- Andy Field. *Discovering statistics using IBM SPSS statistics*. Sage publications limited, 2024.

- Murray J Fisher and Andrea P Marshall. Understanding descriptive statistics. *Australian critical care*, 22(2):93–97, 2009.
- Todd Michael Franke, Timothy Ho, and Christina A Christie. The chi-square test: Often used and more often misinterpreted. *American journal of evaluation*, 33(3):448–458, 2012.
- Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.
- Anastasia Galkin, Kristin Riebe, Ole Streicher, Francois Bonnarel, Mireille Louys, Michèle Sanguillon, Mathieu Servillat, and Markus Nullmeier. Provenance for astrophysical data. In *Provenance and Annotation of Data and Processes: 7th International Provenance and Annotation Workshop, IPAW 2018, London, UK, July 9-10, 2018, Proceedings*, pages 252–256. Springer, 2018.
- Liangcai Gao, Zhi Tang, Xiaofan Lin, Ying Liu, Ruiheng Qiu, and Yongtao Wang. Structure extraction from pdf-based book documents. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 11–20, 2011.
- Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*, 2023.
- Daniel Garijo and Yolanda Gil. A new approach for publishing workflows: abstractions, standards, and linked data. In *Proceedings of the 6th workshop on Workflows in support of large-scale science*, pages 47–56, 2011.
- Daniel Garijo, Maximiliano Osorio, Deborah Khider, Varun Ratnakar, and Yolanda Gil. Okgsoft: An open knowledge graph with machine readable scientific software metadata. In *2019 15th International Conference on eScience (eScience)*, pages 349–358. IEEE, 2019.
- Daniel Garijo Verdejo and Yolanda Gil. Augmenting prov with plans in p-plan: scientific processes as linked data. *CEUR Workshop Proceedings*, 2012.
- Kerstin Gierend, Frank Krüger, Sascha Genehr, Francisca Hartmann, Fabian Siegel, Dagmar Waltemath, Thomas Ganslandt, and Atinkut Alamirrew Zeleke. Capturing provenance information for biomedical data and workflows: A scoping review. 2023.
- Kerstin Gierend, Frank Krüger, Sascha Genehr, Francisca Hartmann, Fabian Siegel, Dagmar Waltemath, Thomas Ganslandt, and Atinkut Alamirrew Zeleke. Provenance information for

- biomedical data and workflows: Scoping review. *Journal of Medical Internet Research*, 26: e51297, 2024.
- Yolanda Gil, Varun Ratnakar, and Daniel Garijo. Ontosoft: Capturing scientific software metadata. In *Proceedings of the 8th International Conference on Knowledge Capture*, pages 1–4, 2015.
- Aidan Gilson, Conrad Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. How well does chatgpt do when taking the medical licensing exams? the implications of large language models for medical education and knowledge assessment. *medRxiv*, pages 2022–12, 2022.
- Louie Giray. Prompt engineering with chatgpt: A guide for academic writers. *Annals of Biomedical Engineering*, pages 1–5, 2023.
- Carole Goble. Position statement: Musings on provenance, workflow and (semantic web) annotations for bioinformatics. In *Workshop on Data Derivation and Provenance, Chicago*, volume 3, 2002.
- Jeremy Goecks, Anton Nekrutenko, James Taylor, and Galaxy Team team@ galaxyproject. org. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86, 2010.
- Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Errell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR, 2023.
- Rafael S Gonçalves and Mark A Musen. The variable quality of metadata about biological samples used in biomedical experiments. *Scientific data*, 6(1):1–15, 2019.
- Peter Goodyear and Lucila Carvalho. Activity centred analysis and design in the evolution of learning networks. In *Proceedings of the 10th International Conference on Networked Learning*, pages 218–225, 2016.
- Peter Goodyear, Lucila Carvalho, and Pippa Yeoman. Activity-centred analysis and design (acad): Core purposes, distinctive qualities and current developments. *Educational Technology Research and Development*, 69:445–464, 2021.
- Todd J Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 31–40, 2007.

- Jane Greenberg. Understanding metadata and metadata schemes. *Cataloging & classification quarterly*, 40(3-4):17–36, 2005.
- Kathleen Gregory, Helena Cousijn, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. Understanding data search as a socio-technical practice. *Journal of Information Science*, 46(4):459–475, 2020.
- Joy M Grossman, Thomas S Bodenheimer, and Kelly McKenzie. Hospital-physician portals: the role of competition in driving clinical data exchange. *Health Affairs*, 25(6):1629–1636, 2006.
- Paul Groth and Luc Moreau. Prov-overview. *W3C Working Group Note*, 1135:881–906, 2013.
- Tobias Grubenmann, Abraham Bernstein, Dmitry Moor, and Sven Seuken. Financing the web of data with delayed-answer auctions. In *Proceedings of the 2018 World Wide Web Conference*, pages 1033–1042, 2018.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
- Thaylon Guedes, Vítor Silva, Marta Mattoso, Marcos VN Bedo, and Daniel de Oliveira. A practical roadmap for provenance capture and data analysis in spark-based scientific workflows. In *2018 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)*, pages 31–41. IEEE, 2018.
- Bernal Jimenez Gutierrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about gpt-3 in-context learning for biomedical ie? think again. *arXiv preprint arXiv:2203.08410*, 2022.
- Deepesh V Hada and Shirish K Shevade. Rexplug: Explainable recommendation using plug-and-play language model. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–91, 2021.
- Olaf Hartig and Jun Zhao. Publishing and consuming provenance metadata on the web of linked data. In *Provenance and Annotation of Data and Processes: Third International Provenance and Annotation Workshop, IPAW 2010, Troy, NY, USA, June 15-16, 2010. Revised Selected Papers 3*, pages 78–90. Springer, 2010.
- James Hendler, Jeanne Holm, Chris Musialek, and George Thomas. Us government linked open data: semantic. data. gov. *IEEE Intelligent Systems*, 27(03):25–31, 2012.

- Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. A survey on provenance: What for? what form? what from? *The VLDB Journal*, 26:881–906, 2017.
- Pieter Heyvaert, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. Merging and enriching dcat feeds to improve discoverability of datasets. In *European Semantic Web Conference*, pages 67–71. Springer, 2015.
- Linda Hickman and Cliff Longman. *Business interviewing*. Addison-Wesley, 1994.
- Jingwei Huang and Mark S Fox. Uncertainty in knowledge provenance. In *European Semantic Web Symposium*, pages 372–387. Springer, 2004.
- Laura D Hughes, Ginger Tsueng, Jack DiGiovanna, Thomas D Horvath, Luke V Rasmussen, Tor C Savidge, Thomas Stoeger, Serdar Turkarslan, Qinglong Wu, Chunlei Wu, et al. Addressing barriers in fair data practices for biomedical data. *Scientific Data*, 10(1):98, 2023.
- Madelon Hulsebos, Wenjing Lin, Shreya Shankar, and Aditya Parameswaran. It took longer than i was expecting: Why is dataset search still so hard? In *Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics*, pages 1–4, 2024.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Trung Dong Huynh and Luc Moreau. Provstore: a public provenance repository. In *International Provenance and Annotation Workshop*, pages 275–277. Springer, 2014.
- Luis-Daniel Ibáñez and Elena Simperl. A comparison of dataset search behaviour of internal versus search engine referred sessions. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 158–168, 2022.
- Nwokedi Idika, Mayank Varia, and Harry Phan. The probabilistic provenance graph. In *2013 IEEE Security and Privacy Workshops*, pages 34–41. IEEE, 2013.
- Matteo Interlandi, Kshitij Shah, Sai Deep Tetali, Muhammad Ali Gulzar, Seunghyun Yoo, Miryung Kim, Todd Millstein, and Tyson Condie. Titian: Data provenance support in spark. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, volume 9, page 216, 2015.
- Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*, 2024.

- ISO/TC 159/SC 4. ISO 9241-210:2010 ergonomics of human-system interaction – part 210: Human-centred design for interactive systems. <https://www.iso.org/standard/52075.html>, 2010. International Organization for Standardization.
- Lakshmi S Iyer, Babita Gupta, and Nakul Johri. Performance, scalability and reliability issues in web applications. *Industrial Management & Data Systems*, 105(5):561–576, 2005.
- Annika Jacobsen, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T Evelo, et al. Fair principles: interpretations and implementation considerations, 2020.
- Ng Chirk Jenn. Designing a questionnaire. *Malaysian family physician: the official journal of the Academy of Family Physicians of Malaysia*, 1(1):32, 2006.
- Neil Jethani, Simon Jones, Nicholas Genes, Vincent J Major, Ian S Jaffe, Anthony B Cardillo, Noah Heilenbach, Nadia Fazal Ali, Luke J Bonanni, Andrew J Clayburn, et al. Evaluating chatgpt in information extraction: a case study of extracting cognitive exam dates and scores. 2023.
- Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 2023.
- Prasad Jogalekar and Murray Woodside. Evaluating the scalability of distributed systems. *IEEE Transactions on parallel and distributed systems*, 11(6):589–603, 2000.
- Marco Johns, Thierry Meurers, Felix N Wirth, Anna C Haber, Armin Mueller, Mehmed Halilovic, Felix Balzer, and Fabian Prasser. Data provenance in biomedical research: scoping review. *Journal of Medical Internet Research*, 25:e42289, 2023.
- Rex E Jung, Rane A Flores, and Dan Hunter. A new measure of imagination ability: Anatomical brain imaging correlates. *Frontiers in psychology*, 7:496, 2016.
- Emilia Kacprzak, Laura Koesten, Jeni Tennison, and Elena Simperl. Characterising dataset search queries. In *Companion Proceedings of the The Web Conference 2018*, pages 1485–1488, 2018.
- Uday Kamath, Kevin Keenan, Garrett Somers, and Sarah Sorenson. Llm challenges and solutions. In *Large Language Models: A Deep Dive: Bridging Theory and Practice*, pages 219–274. Springer, 2024.
- Hyun Kang. Sample size determination and power analysis using the g* power software. *Journal of educational evaluation for health professions*, 18, 2021.

- Swati Kanwal. *Exploring the potential of LLMs for biomedical relation extraction*. PhD thesis, University of British Columbia, 2024.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.
- Maxat Kassen. A promising phenomenon of open data: A case study of the chicago open data project. *Government information quarterly*, 30(4):508–513, 2013.
- Makoto P Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. A test collection for ad-hoc dataset retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2450–2456, 2021.
- Saadiq Rauf Khan, Vinit Chandak, and Sougata Mukherjea. Evaluating llms for visualization generation and understanding. *Discover Data*, 3(1):15, 2025.
- Hae-Young Kim. Analysis of variance (anova) comparing means of more than two groups. *Restorative dentistry & endodontics*, 39(1):74–77, 2014.
- Barbara Kitchenham et al. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.
- Ryan KL Ko and Mark A Will. Progger: An efficient, tamper-evident kernel-space logger for cloud data provenance tracking. In *2014 IEEE 7th International Conference on Cloud Computing*, pages 881–889. IEEE, 2014.
- Mateja Kocbek, Gregor Jošt, Marjan Heričko, and Gregor Polančič. Business process model and notation: The current state of affairs. *Computer Science and Information Systems*, 12(2): 509–539, 2015.
- Neesha Kodagoda, Sheila Pontis, Donal Simmie, Simon Attfield, BL William Wong, Ann Blandford, and Chris Hankin. Using machine learning to infer reasoning provenance from user interaction log data: based on the data/frame theory of sensemaking. *Journal of Cognitive Engineering and Decision Making*, 11(1):23–41, 2017.
- L. Koesten. *A User Centred Perspective on Structured Data Discovery*. University of Southampton, Faculty of Engineering and Physical Sciences, PhD Thesis, [pagination], 2019a.
- Laura Koesten. *A user centred perspective on structured data discovery*. PhD thesis, University of Southampton, 2019b.

- Laura Koesten, Elena Simperl, Tom Blount, Emilia Kacprzak, and Jeni Tennison. Everything you always wanted to know about a dataset: Studies in data summarisation. *International Journal of Human-Computer Studies*, 135:102367, 2020.
- Troy Kohwalter, Thiago Oliveira, Juliana Freire, Esteban Clua, and Leonardo Murta. Prov viewer: A graph-based visualization tool for interactive ex-ploration of provenance data. In *Provenance and Annotation of Data and Processes: 6th International Provenance and Annotation Workshop, IPAW 2016, McLean, VA, USA, June 7-8, 2016, Proceedings 6*, pages 71–82. Springer, 2016.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- B Koning. Extracting sections from pdf-formatted cti reports. B.S. thesis, University of Twente, 2022.
- Phillip S Kott. A model-based look at linear regression with survey data. *The american statistician*, 45(2):107–112, 1991.
- Christos Koutras, George Siachamis, Andra Ionescu, Kyriakos Psarakis, Jerry Brons, Marios Fragkoulis, Christoph Lofi, Angela Bonifati, and Asterios Katsifodimos. Valentine: Evaluating matching techniques for dataset discovery. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 468–479. IEEE, 2021.
- Thomas Krämer, Andrea Papenmeier, Zeljko Carevic, Dagmar Kern, and Brigitte Mathiak. Data-seeking behaviour in the social sciences. *International Journal on Digital Libraries*, 22(2):175–195, 2021.
- Sven R Kunze and Sören Auer. Dataset retrieval. In *2013 IEEE Seventh International Conference on Semantic Computing*, pages 1–8. IEEE, 2013.
- Tzu-Lin Kuo, Tzu-Wei Chiu, Tzung-Sheng Lin, Sheng-Yang Wu, Chao-Wei Huang, and Yun-Nung Chen. A survey of generative information retrieval. *arXiv preprint arXiv:2406.01197*, 2024.
- Yrjo Lappalainen and Nikesh Narayanan. Aisha: A custom ai library chatbot using the chatgpt api. *Journal of Web Librarianship*, 17(3):37–58, 2023.
- Luca Lauro, Pasquale Leonardo Lazzaro, Marialaura Lazzaro, Paolo Missier, and Riccardo Torlone. An llm-guided platform for multi-granular collection and management of data provenance. 2024.

- Alexa Lazard and Lucy Atkinson. Putting the visual back in visual communication. *International Journal of Information Management*, 37(3):1449–1458, 2017. .
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- Jeremy Leipzig, Daniel Nüst, Charles Tapley Hoyt, Karthik Ram, and Jane Greenberg. The role of metadata in reproducible computational research. *Patterns*, 2(9), 2021.
- Simone Leo, Michael R Crusoe, Laura Rodríguez-Navas, Raül Sirvent, Alexander Kanitz, Paul De Geest, Rudolf Wittner, Luca Pireddu, Daniel Garijo, José M Fernández, et al. Recording provenance of workflow runs with ro-crate. *PLOS One*, 19(9):e0309210, 2024.
- Barbara Lerner and Emery Boose. {RDataTracker}: Collecting provenance in an interactive scripting environment. In *6th USENIX Workshop on the Theory and Practice of Provenance (TaPP 2014)*, 2014.
- Dirk Lewandowski. Vertical search. In *Understanding Search Engines*, pages 119–136. Springer, 2023.
- Mike Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.
- Mingchen Li, Ming Chen, Huixue Zhou, and Rui Zhang. Petailor: Improving large language model by tailored chunk scorer in biomedical triple extraction. *arXiv preprint arXiv:2310.18463*, 2023.
- Catherine Compton Lilly. Book review: Creswell, john.(1997). qualitative inquiry and re-search design: Choosing among five traditions. *Networks: An Online Journal for Teacher Research*, 1(1):62–62, 1998.
- Bin Liu and HV Jagadish. Datalens: making a good first impression. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 1115–1118, 2009.
- Chang Liu, Matthew Kim, Michael Rueschman, and Satya S Sahoo. Provcare: A large-scale semantic provenance resource for scientific reproducibility. In *Provenance in Data Science: From Data Models to Context-Aware Knowledge Graphs*, pages 59–73. Springer, 2020.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

- Jianzheng Liu, Jie Li, Weifeng Li, and Jiansheng Wu. Rethinking big data: A review on the data quality and usage issues. *ISPRS journal of photogrammetry and remote sensing*, 115: 134–142, 2016.
- Qi Liu, Yongyi He, Tong Xu, Defu Lian, Che Liu, Zhi Zheng, and Enhong Chen. Unimel: A unified framework for multimodal entity linking with large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1909–1919, 2024b.
- Yilun Liu, Shimin Tao, Weibin Meng, Jingyu Wang, Wenbing Ma, Yanqing Zhao, Yuhang Chen, Hao Yang, Yanfei Jiang, and Xun Chen. Logprompt: Prompt engineering towards zero-shot and interpretable log analysis. *arXiv preprint arXiv:2308.07610*, 2023.
- Felicitas Löffler, Valentin Wesp, Birgitta König-Ries, and Friederike Klan. Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs? *PloS one*, 16(3):e0246099, 2021.
- Felicitas Löffler, Fateme Shafiei, René Witte, Birgitta König-Ries, and Friederike Klan. Semantic search for biological datasets: A usability study on modes of querying and explaining search results. In *BTW 2023*, pages 851–864. Gesellschaft für Informatik eV, 2023.
- Fadi Maali, John Erickson, and Phil Archer. Data catalog vocabulary (dcat). w3c recommendation. *The World Wide Web Consortium*, 2014.
- Allan J MacKenzie-Graham, Arash Payan, Ivo D Dinov, John D Van Horn, and Arthur W Toga. Neuroimaging data provenance using the Ioni pipeline workflow environment. In *Provenance and Annotation of Data and Processes: Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008. Revised Selected Papers 2*, pages 208–220. Springer, 2008.
- Peter Macko and Marc Chiarini. Collecting provenance via the xen hypervisor. In *3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP 11)*, 2011.
- Sara Magliacane. Reconstructing provenance. In *International Semantic Web Conference*, pages 399–406. Springer, 2012.
- Christopher D Manning. *Introduction to information retrieval*. Syngress Publishing., 2008.
- Gary Marchionini, Stephanie W Haas, Junliang Zhang, and Jonathan Elsas. Accessing government statistical information. *Computer*, 38(12):52–61, 2005.

- Laura Haak Marcial and Bradley M Hemminger. Scientific data repositories on the web: An initial survey. *Journal of the American Society for Information Science and Technology*, 61(10):2029–2048, 2010.
- Keith D Markman, William MP Klein, and Julie A Suhr. *Handbook of imagination and mental simulation*. Psychology Press, 2012.
- Milan Markovic, Peter Edwards, and David Corsar. Sc-prov: A provenance vocabulary for social computation. In *Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers 5*, pages 285–287. Springer, 2015.
- Deborah L Martin, Jennifer L Hoff, Roger A Gard, Richard J Gregosky, Hobert W Jones, Cheryl A Kirkwood, Donald G Morris, Tracey E Shinsato, and Cheryl L Willott-Moore. Data collection, processing, validation, and verification. *Health physics*, 95(1):36–46, 2008.
- Matthew S Mayernik. Metadata. *KO Knowledge Organization*, 47(8):696–713, 2021.
- Susanne Mayr, Edgar Erdfelder, Axel Buchner, and Franz Faul. A short tutorial of gpower. *Tutorials in quantitative methods for psychology*, 3(2):51–59, 2007.
- Richard McClatchey, Jetendr Shamdasani, Andrew Branson, and Kamran Munir. Provenance support for medical research. In *Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers 5*, pages 291–293. Springer, 2015.
- James P McCusker and Deborah L McGuinness. Explorations into the provenance of high throughput biomedical experiments. In *Provenance and Annotation of Data and Processes: Third International Provenance and Annotation Workshop, IPAW 2010, Troy, NY, USA, June 15-16, 2010. Revised Selected Papers 3*, pages 120–128. Springer, 2010a.
- Jamie McCusker and D McGuinness. Provenance of high throughput biomedical experiments. In *International Provenance and Annotations Workshop: 15-16 June 2010; Troy, NY*. Rensselaer Polytechnic University Troy, NY, 2010b.
- Michael McDonnell and Ali Shiri. Social search: A taxonomy of, and a user-centred approach to, social web search. *Program*, 45(1):6–28, 2011.
- Timothy McPhillips, Shawn Bowers, Khalid Belhajjame, and Bertram Ludäscher. Retrospective provenance without a runtime provenance recorder. In *7th USENIX Workshop on the Theory and Practice of Provenance (TaPP 15)*, 2015.

- Jeffrey Mervis. Agencies rally to tackle big data, 2012.
- Bertalan Meskó. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of Medical Internet Research*, 25:e50638, 2023.
- Seifeddine Messaoud, Abbas Bradai, Syed Hashim Raza Bukhari, Pham Tran Anh Quang, Olfa Ben Ahmed, and Mohamed Atri. A survey on machine learning in internet of things: Algorithms, strategies, and applications. *Internet of Things*, 12:100314, 2020.
- Peter Mika. Flink: Semantic web technology for the extraction and analysis of social networks. *Journal of web semantics*, 3(2-3):211–223, 2005.
- Milan Milanovic, Dragan Gašević, and Gerd Wagner. Combining rules and activities for modeling service-based business processes. In *2008 12th Enterprise Distributed Object Computing Conference Workshops*, pages 11–22. IEEE, 2008.
- Simon Miles, Paul Groth, Miguel Branco, and Luc Moreau. The requirements of using provenance in e-science experiments. *Journal of Grid Computing*, 5:1–25, 2007.
- Beverley C Millar and Michelle Lim. The role of visual abstracts in the dissemination of medical research. *The Ulster medical journal*, 91(2):67, 2022.
- Paolo Missier, Khalid Belhajjame, and James Cheney. The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 773–776, 2013a.
- Paolo Missier, Saumen Dey, Khalid Belhajjame, Víctor Cuevas-Vicentín, and Bertram Ludäscher. {D-PROV}: Extending the {PROV} provenance model with {Workflow} structure. In *5th USENIX Workshop on the Theory and Practice of Provenance (TaPP 13)*, 2013b.
- Paolo Missier, Fernando Chirigati, Yaxing Wei, David Koop, and Saumen Dey. Provenance storage, querying, and visualization in pbase. In *International Provenance and Annotation Workshop (IPAW)*, volume 8628, page 239. Springer, 2015.
- Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, et al. The open provenance model core specification (v1. 1). *Future generation computer systems*, 27(6):743–756, 2011.
- Luc Moreau, Paul Groth, James Cheney, Timothy Lebo, and Simon Miles. The rationale of prov. *Journal of Web Semantics*, 35:235–257, 2015.
- Kiran-Kumar Muniswamy-Reddy and Margo Seltzer. Provenance as first class cloud data. *ACM SIGOPS Operating Systems Review*, 43(4):11–16, 2010.

- Leonardo Murta, Vanessa Braganholo, Fernando Chirigati, David Koop, and Juliana Freire. noworkflow: capturing and analyzing provenance of scripts. In *Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers 5*, pages 71–83. Springer, 2015.
- Mehwish Nasir. A survey of software estimation techniques and project planning practices. In *Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD'06)*, pages 305–310. IEEE, 2006.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.
- Phuc Nguyen, Kazutoshi Shinoda, Taku Sakamoto, Diana Andreea Petrescu, Hung Nghiep Tran, Atsuhiko Takasu, Akiko Aizawa, and Hideaki Takeda. Nii table linker at the ntcir-15 data search task. In *Proceedings of the NTCIR-15 Conference*, 2020.
- Thanh Tam Nguyen, Quoc Viet Hung Nguyen, Matthias Weidlich, and Karl Aberer. Result selection and summarization for web table search. In *2015 IEEE 31st International Conference on Data Engineering*, pages 231–242. IEEE, 2015.
- Isak Nordgren and Gustaf E Svensson. Prompt engineering and its usability to improve modern psychology chatbots, 2023.
- Lucila Ohno-Machado, Susanna-Assunta Sansone, George Alter, Ian Fore, Jeffrey Grethe, Hua Xu, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, Anupama E Gururaj, Elizabeth Bell, et al. Finding useful data across multiple biomedical data repositories using datamed. *Nature genetics*, 49(6):816–819, 2017.
- Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R Pocock, Anil Wipat, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, 2004.
- Kingsley Okoye and Samira Hosseini. T-test statistics in r: Independent samples, paired sample, and one sample t-tests. In *R Programming: Statistical Data Analysis in Research*, pages 159–186. Springer, 2024.
- Kian Win Ong, Yannis Papakonstantinou, and Romain Vernoux. The sql++ query language: Configurable, unifying and semi-structured. *arXiv preprint arXiv:1405.3631*, 2014.

- OpenAI. Introducing gpt-4o, 2024. URL <https://openai.com/index/hello-gpt-4o/>. Accessed: 2025-08-03.
- M Tamer Özsu. A survey of rdf data management systems. *Frontiers of Computer Science*, 10: 418–432, 2016.
- Michelle O’reilly and Nicola Parker. ‘unsatisfactory saturation’: a critical exploration of the notion of saturated sample sizes in qualitative research. *Qualitative research*, 13(2):190–197, 2013.
- Bofeng Pan, Natalia Stakhanova, and Suprio Ray. Data provenance in security and privacy. *ACM Computing Surveys*, 2023.
- Aaron Parisi, Yao Zhao, and Noah Fiedel. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*, 2022.
- Carlos Luis Parra-Calderón, Ferran Sanz, and Leslie D McIntosh. The challenge of the effective implementation of fair principles in biomedical research. *Methods of Information in Medicine*, 59(04/05):117–118, 2020.
- Nikolaus Nova Parulian and Bertram Ludäscher. Trust the process: Analyzing prospective provenance for data cleaning. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1513–1523, 2023.
- Thomas Pasquier, Xueyuan Han, Mark Goldstein, Thomas Moyer, David Eysers, Margo Seltzer, and Jean Bacon. Practical whole-system provenance capture. In *Proceedings of the 2017 Symposium on Cloud Computing*, pages 405–418, 2017.
- Thomas Pasquier, Xueyuan Han, Thomas Moyer, Adam Bates, Olivier Hermant, David Eysers, Jean Bacon, and Margo Seltzer. Runtime analysis of whole-system provenance. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 1601–1616, 2018.
- Norman W Paton, Jiaoyan Chen, and Zhenyu Wu. Dataset discovery and exploration: A survey. *ACM Computing Surveys*, 56(4):1–37, 2023.
- Braja Gopal Patra, Kirk Roberts, and Hulin Wu. A content-based dataset recommendation system for researchers—a case study on gene expression omnibus (geo) repository. *Database*, 2020, 2020.
- Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

- Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*, 2023.
- Wenjun Peng, Guiyang Li, Yue Jiang, Zilong Wang, Dan Ou, Xiaoyi Zeng, Derong Xu, Tong Xu, and Enhong Chen. Large language model based long-tail query rewriting in taobao search. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 20–28, 2024.
- Reuben Pengelly, Babatunde Rowaiye, Karen Pickard, Brendan J. Moran, Sanjeev Dayal, William Tapper, Alexander Mirnezami, Tom Cecil, Faheez Mohamed, Norman Carr, and Sarah Ennis. Analysis of mutation and loss of heterozygosity by whole-exome sequencing yields insights into pseudomyxoma peritonei. *Annals of Oncology*, 29(3):725–733, 2018.
- Rebekah Penrice-Randal, Xiaofeng Dong, Andrew George Shapanis, Aaron Gardner, Nicholas Harding, Jelmer Legebeke, Jenny Lord, Andres F Vallejo, Stephen Poole, Nathan J Brendish, Catherine Hartley, Anthony P Williams, Gabrielle Wheway, Marta E Polak, Fabio Strazzeri, James P R Schofield, Paul J Skipp, Julian A Hiscox, Tristan W Clark, and Diana Baralle. Blood gene expression predicts intensive care unit admission in hospitalized patients with covid-19. *Scientific Reports*, 12:578, 2022.
- Felipe Pezoa, Juan L Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoč. Foundations of json schema. In *Proceedings of the 25th international conference on World Wide Web*, pages 263–273, 2016.
- Rakesh Pimplikar and Sunita Sarawagi. Answering table queries on the web using column keywords. *arXiv preprint arXiv:1207.0132*, 2012.
- Débora Pina, Adriane Chapman, Liliane Kunstmann, Daniel de Oliveira, and Marta Mattoso. Dlprov: A data-centric support for deep learning workflow analyses. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*, pages 77–85, 2024.
- Peter Pirolli and Ramana Rao. Table lens as a tool for making sense of data. In *Proceedings of the workshop on Advanced visual interfaces*, pages 67–80, 1996.
- Adam Pocock. Tribuo: Machine learning with provenance in java. *arXiv preprint arXiv:2110.03022*, 2021.
- Maciej P Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1): 1569, 2024.

- Fina Polat, Ilaria Tiddi, and Paul Groth. Testing prompt engineering methods for knowledge extraction from text. *Semantic Web. Under Review*, 2024.
- Thomas P Quinn, Ionas Erb, Mark F Richardson, and Tamsyn M Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878, 2018.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.
- Sudha Ram and Jun Liu. Understanding the semantics of data provenance to support active conceptual modeling. In *Active Conceptual Modeling of Learning: Next Generation Learning-Base System Development 1*, pages 17–29. Springer, 2007.
- Sudha Ram, Jun Liu, et al. A new perspective on semantics of data provenance. *SWPM*, 526, 2009.
- Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy, and Gully APC Burns. Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, 7: 1–10, 2012.
- Torben Sølbeck Rasmussen, Caroline Märta Junker Mentzel, Witold Kot, Josué Leonardo Castro-Mejía, Simone Zuffa, Jonathan Richard Swann, Lars Hestbjerg Hansen, Finn Kvist Vogensen, Axel Kornerup Hansen, and Dennis Sandris Nielsen. Faecal virome transplantation decreases symptoms of type 2 diabetes and obesity in a murine model. *Gut*, 69:1975–1985, 2020. .
- Melanie Revilla and Jan Karem Höhne. How long do respondents think online surveys should be? new evidence from two online panels in germany. *International Journal of Market Research*, 62(5):538–545, 2020.
- Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.
- Jenn Riley. Understanding metadata. *Washington DC, United States: National Information Standards Organization (<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>)*, 23:7–10, 2017.

- Kirk Roberts, Anupama E Gururaj, Xiaoling Chen, Saeid Pournejati, William R Hersh, Dina Demner-Fushman, Lucila Ohno-Machado, Trevor Cohen, and Hua Xu. Information retrieval for biomedical datasets: the 2016 biocaddie dataset retrieval challenge. *Database*, 2017, 2017.
- Colin Robson and Kieran McCartan. *Real world research*. Wiley Global Education, 2016.
- Philippe Rocca-Serra, Marco Brandizi, Eamonn Maguire, Nataliya Sklyar, Chris Taylor, Kimberly Begley, Dawn Field, Stephen Harris, Winston Hide, Oliver Hofmann, et al. Isa software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18):2354–2356, 2010.
- Margarita Ruiz-Olazar, Evandro S Rocha, Sueli S Rabaça, Carlos Eduardo Ribas, Amanda S Nascimento, and Kelly R Braghetto. A review of guidelines and models for representation of provenance information from neuroscience experiments. In *Provenance and Annotation of Data and Processes: 6th International Provenance and Annotation Workshop, IPAW 2016, McLean, VA, USA, June 7-8, 2016, Proceedings 6*, pages 222–225. Springer, 2016.
- Carlos Sáenz-Adán, Luc Moreau, Beatriz Pérez, Simon Miles, and Francisco J García-Izquierdo. Automating provenance capture in software engineering with uml2prov. In *Provenance and Annotation of Data and Processes: 7th International Provenance and Annotation Workshop, IPAW 2018, London, UK, July 9-10, 2018, Proceedings*, pages 58–70. Springer, 2018.
- Satya S Sahoo, Joshua Valdez, Matthew Kim, Michael Rueschman, and Susan Redline. Provcare: characterizing scientific reproducibility of biomedical research studies using semantic provenance metadata. *International journal of medical informatics*, 121:10–18, 2019.
- Satya S Sahoo, Matthew D Turner, Lei Wang, Jose Luis Ambite, Abhishek Appaji, Arcot Rajasekar, Howard M Lander, Yue Wang, and Jessica A Turner. Neurobridge ontology: computable provenance metadata to give the long tail of neuroimaging data a fair chance for secondary use. *Frontiers in Neuroinformatics*, 17:1216443, 2023.
- Satya S Sahoo, Joseph M Plasek, Hua Xu, Özlem Uzuner, Trevor Cohen, Meliha Yetisgen, Hongfang Liu, Stéphane Meystre, and Yanshan Wang. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. *Journal of the American Medical Informatics Association*, page ocae074, 2024.
- Gerard Salton, Edward A Fox, and Harry Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983.
- Sheeba Samuel. *A provenance-based semantic approach to support understandability, reproducibility, and reuse of scientific experiments*. PhD thesis, 2019.

- Sheeba Samuel and Birgitta König-Ries. Reproduce-me: ontology-based data access for reproducibility of microscopy experiments. In *The Semantic Web: ESWC 2017 Satellite Events: ESWC 2017 Satellite Events, Portorož, Slovenia, May 28–June 1, 2017, Revised Selected Papers 14*, pages 17–20. Springer, 2017.
- Sheeba Samuel and Birgitta König-Ries. End-to-end provenance representation for the understandability and reproducibility of scientific experiments using a semantic approach. *Journal of biomedical semantics*, 13(1):1, 2022.
- Margarete Sandelowski. Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies. *Research in nursing & health*, 23(3):246–255, 2000.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.
- Susanna-Assunta Sansone, Philippe Rocca-Serra, Marco Brandizi, Alvis Brazma, Dawn Field, Jennifer Fostel, Andrew G Garrow, Jack Gilbert, Federico Goodsaid, Nigel Hardy, et al. The first rsbi (isa-tab) workshop: “can a simple format work for complex studies?”. *OMICS A Journal of Integrative Biology*, 12(2):143–149, 2008.
- Susanna-Assunta Sansone, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, George Alter, Jeffrey S Grethe, Hua Xu, Ian M Fore, Jared Lyle, Anupama E Gururaj, Xiaoling Chen, et al. Dats, the data tag suite to enable discoverability of datasets. *Scientific data*, 4(1):1–8, 2017.
- Ugis Sarkans, Mikhail Gostev, Awais Athar, Ehsan Behrangi, Olga Melnichuk, Ahmed Ali, Jasmine Minguet, Juan Camillo Rada, Catherine Snow, Andrew Tikhonov, et al. The biostudies database—one stop shop for all data supporting a life sciences study. *Nucleic acids research*, 46(D1):D1266–D1270, 2018.
- Mark Saunders, Philip Lewis, and Adrian Thornhill. *Research methods for business students*. Pearson education, 2009.
- Timo Schick and Hinrich Schütze. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, 2021.
- Douglas C Schmidt, Jesse Spencer-Smith, Quchen Fu, and Jules White. Towards a catalog of prompt patterns to enhance the discipline of prompt engineering. *ACM SIGAda Ada Letters*, 43(2):43–51, 2024.

- Oladipupo A Sennaike, Mohammad Waqar, Edobor Osagie, Islam Hassan, Arkadiusz Stasiewicz, Lukasz Porwol, and Adegboyega Ojo. Towards intelligent open data platforms: Discovering relatedness in datasets. In *2017 Intelligent Systems Conference (IntelliSys)*, pages 414–421. IEEE, 2017.
- B Séroussi et al. The common provenance model: Capturing distributed provenance in life sciences processes. *Challenges of Trustable AI and Added-Value on Health: Proceed-ings of MIE 2022*, 294:415, 2022.
- H. Sharp, J. Preece, and Y. Rogers. *Interaction Design: Beyond Human-Computer Interaction*. Wiley, 2019. ISBN 9781119547259. URL <https://books.google.co.uk/books?id=HreODwAAQBAJ>.
- Junxiao Shen, John J Dudley, Jingyao Zheng, Bill Byrne, and Per Ola Kristensson. Promptor: A conversational and autonomous prompt generation agent for intelligent text entry techniques. *arXiv preprint arXiv:2310.08101*, 2023.
- Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Auto-prompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.
- Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*, 2022.
- Katharina Sielemann, Alenka Hafner, and Boas Pucker. The reuse of public datasets in the life sciences: potential risks and rewards. *PeerJ*, 8:e9954, 2020.
- Vítor Silva, Renan Souza, Jose Camata, Daniel de Oliveira, Patrick Valduriez, Alvaro LGA Coutinho, and Marta Mattoso. Capturing provenance for runtime data analysis in computational science and engineering applications. In *Provenance and Annotation of Data and Processes: 7th International Provenance and Annotation Workshop, IPAW 2018, London, UK, July 9-10, 2018, Proceedings*, pages 183–187. Springer, 2018.
- Yogesh L Simmhan, Beth Plale, and Dennis Gannon. A survey of data provenance in e-science. *ACM Sigmod Record*, 34(3):31–36, 2005.
- Ayush Singhal and Jaideep Srivastava. Research dataset discovery from research publications using web context. In *Web Intelligence*, volume 15, pages 81–99. IOS Press, 2017.
- Anshu Sinha, George Hripcsak, and Marianthi Markatou. Large datasets in biomedicine: a discussion of salient analytic issues. *Journal of the American Medical Informatics Association*, 16(6):759–767, 2009.

- Shanmugasundaram Sivakumar. Performance optimization of large language models (llms) in web applications. 2024.
- Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, et al. Packaging research artefacts with ro-crate. *Data Science*, 5(2):97–138, 2022.
- Jin Hyun Son and Myoung Ho Kim. Improving the performance of time-constrained workflow processing. *Journal of Systems and Software*, 58(3):211–219, 2001.
- Katrina Sostek, Daniel M Russell, Nitesh Goyal, Tarfah Alrashed, Stella Dugall, and Natasha Noy. Discovering datasets on the web scale: Challenges and recommendations for google dataset search. *Harvard Data Science Review*, (Special Issue 4), 2024.
- Manolis Stamatogiannakis, Elias Athanasopoulos, Herbert Bos, and Paul Groth. Prov 2r: practical provenance analysis of unstructured processes. *ACM Transactions on Internet Technology (TOIT)*, 17(4):1–24, 2017.
- Salmin Sultana and Elisa Bertino. A comprehensive model for provenance. In *Advances in Conceptual Modeling: ER 2012 Workshops CMS, ECDM-NoCoDA, MoDIC, MORE-BI, RIGiM, SeCoGIS, WISM, Florence, Italy, October 15-18, 2012. Proceedings 31*, pages 121–130. Springer, 2012.
- Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint arXiv:2303.07992*, 2023.
- Yiyi Tang, Ziyang Xiao, Xue Li, Qingpeng Zhang, Esther WY Chan, Ian CK Wong, and Research Data Collaboration Task Force. Large language model in medical information extraction from titles and abstracts with prompt engineering strategies: A comparative study of gpt-3.5 and gpt-4. *medRxiv*, pages 2024–03, 2024.
- Mohsen Tavakol and Reg Dennick. Making sense of cronbach’s alpha. *International journal of medical education*, 2:53, 2011.
- Pittawat Taveekitworachai and Ruck Thawonmas. Enhancing novelty in chatgpt responses: Incorporating random word brainstorming. In *Proceedings of the 13th International Conference on Advances in Information Technology*, pages 1–7, 2023.
- Douglas Teodoro, Luc Mottin, Julien Gobeill, Arnaud Gaudinat, Thérèse Vachon, and Patrick Ruch. Improving average ranking precision in user searches for biomedical research datasets. *Database*, 2017:bax083, 2017.

- Surendrabikram Thapa and Surabhi Adhikari. Chatgpt, bard, and large language models for biomedical research: opportunities and pitfalls. *Annals of biomedical engineering*, 51(12): 2647–2651, 2023.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.
- Jennifer L Thøgersen and Pia Borlund. Researcher attitudes toward data sharing in public data repositories: a meta-evaluation of studies on researcher data sharing. *Journal of Documentation*, 78(7):1–17, 2022.
- Paul Thomas, Rollin Omari, and Tom Rowlands. Towards searching amongst tables. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–4, 2015.
- Jeffrey Thorsby, Genie NL Stowers, Kristen Wolslegel, and Ellie Tumbuan. Understanding the content and features of open data portals in american cities. *Government information quarterly*, 34(1):53–61, 2017.
- Leonardo Thurler, Sidney Melo, Leonardo Murta, Troy Kohwalter, and Esteban Clua. Using provenance and replay for qualitative analysis of gameplay sessions. *Entertainment Computing*, 52:100778, 2025.
- Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1): bbad493, 2024.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Ginger Tsueng, Marco A Alvarado Cano, José Bento, Candice Czech, Mengjia Kang, Lars Pache, Luke V Rasmussen, Tor C Savidge, Justin Starren, Qinglong Wu, et al. Developing a standardized but extendable framework to increase the findability of infectious disease datasets. *Scientific Data*, 10(1):99, 2023.
- Hannes Ulrich, Ann-Kristin Kock-Schoppenhauer, Noemi Deppenwiese, Robert Gött, Jori Kern, Martin Lablans, Raphael W Majeed, Mark R Stöhr, Jürgen Stausberg, Julian Varghese, et al. Understanding the nature of metadata: systematic review. *Journal of medical Internet research*, 24(1):e25440, 2022.

- Joshua Valdez, Matthew Kim, Michael Rueschman, Vimig Socrates, Susan Redline, and Satya S Sahoo. Provcare semantic provenance knowledgebase: evaluating scientific reproducibility of research studies. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1705. American Medical Informatics Association, 2017.
- Arnold POS Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries*, pages 521–530, 2010.
- Aishwarya Vijayan. A prompt engineering approach for structured data extraction from unstructured text using conversational llms. In *Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence*, pages 183–189, 2023.
- Mark Von Rosing, Stephen White, Fred Cummins, and Henk De Man. Business process model and notation-bpmn., 2015.
- Naomi Waithira, Mavuto Mukaka, Evelyne Kestelyn, Keitcheya Chotthanawathit, Dung Nguyen Thi Phuong, Hoa Nguyen Thanh, Anne Osterrieder, Trudie Lang, and Phaik Yeong Cheah. Data sharing and reuse in clinical research: Are we there yet? a cross-sectional study on progress, challenges and opportunities in Imics. *PLOS Global Public Health*, 4(11):e0003392, 2024.
- Alexander M Waldrop, John B Cheadle, Kira Bradford, Alexander Preiss, Robert Chew, Jonathan R Holt, Nathan Braswell, Matt Watson, Andrew Crerar, Chris M Ball, et al. Dug: A semantic search engine leveraging peer-reviewed knowledge to span biomedical data repositories. *bioRxiv*, pages 2021–07, 2021.
- Alexander M Waldrop, John B Cheadle, Kira Bradford, Alexander Preiss, Robert Chew, Jonathan R Holt, Yaphet Kebede, Nathan Braswell, Matt Watson, Virginia Hench, et al. Dug: a semantic search engine leveraging peer-reviewed knowledge to query biomedical data repositories. *Bioinformatics*, 38(12):3252–3258, 2022.
- Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019a.

- Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022.
- Dixuan Wang, Yanda Li, Junyuan Jiang, Zepeng Ding, Guochao Jiang, Jiaqing Liang, and Deqing Yang. Tokenization matters! degrading large language models through challenging their tokenization. *arXiv preprint arXiv:2405.17067*, 2024.
- Shuyue Wang and P Jin. A brief summary of prompting in using gpt models. 2023.
- Xiaxia Wang, Jinchi Chen, Shuxin Li, Gong Cheng, Jeff Z Pan, Evgeny Kharlamov, and Yuzhong Qu. A framework for evaluating snippet generation for dataset search. In *International Semantic Web Conference*, pages 680–697. Springer, 2019b.
- Zichen Wang, Alexander Lachmann, and Avi Ma’ayan. Mining data and metadata from the gene expression omnibus. *Biophysical reviews*, 11:103–110, 2019c.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b.
- Wei Wei, Zhanglong Ji, Yupeng He, Kai Zhang, Yuanchi Ha, Qi Li, and Lucila Ohno-Machado. Finding relevant biomedical datasets: the uc san diego solution for the biocaddie retrieval challenge. *Database*, 2018: bay017, 2018.
- Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.
- Andrea Wiggins, Alyson Young, and Melissa A Kenney. Exploring visual representations to support data re-use for interdisciplinary science. *Proceedings of the Association for Information Science and Technology*, 55(1):554–563, 2018.

- Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.
- Lloyd G Williams and Connie U Smith. Qsemsm: Quantitative scalability evaluation method. In *Proc. CMG*, 2005.
- Cerys Willoughby and Jeremy G Frey. Documentation and visualisation of workflows for effective communication, collaboration and publication@ source. *International Journal of Digital Curation*, 12(1):72–87, 2017.
- Rudolf Wittner, Petr Holub, Heimo Müller, Joerg Geiger, Carole Goble, Stian Soiland-Reyes, Luca Pireddu, Francesca Frexia, Cecilia Mascia, Elliot Fairweather, et al. Iso 23494: Biotechnology–provenance information model for biological specimen and data. In *International Provenance and Annotation Workshop*, pages 222–225. Springer, 2020.
- Rudolf Wittner, Petr Holub, Heimo Müller, Joerg Geiger, Carole Goble, Stian Soiland-Reyes, Luca Pireddu, Francesca Frexia, Cecilia Mascia, Elliot Fairweather, et al. Iso 23494: Biotechnology–provenance information model for biological specimen and data. In *Provenance and Annotation of Data and Processes: 8th and 9th International Provenance and Annotation Workshop, IPAW 2020+ IPAW 2021, Virtual Event, July 19–22, 2021, Proceedings*, pages 222–225. Springer, 2021.
- Rudolf Wittner, Matej Gallo, Francesca Frexia, Simone Leo, Luca Pireddu, Cecilia Mascia, Markus Plass, Stian Soiland-Reyes, Heimo Müller, Jörg Geiger, et al. Linking provenance and its metadata in multi-organizational environments. 2023.
- Rudolf Wittner, Petr Holub, Cecilia Mascia, Francesca Frexia, Heimo Müller, Markus Plass, Clare Allocca, Fay Betsou, Tony Burdett, Ibon Cancio, et al. Toward a common standard for data and specimen provenance in life sciences. *Learning health systems*, 8(1):e10365, 2024a.
- Rudolf Wittner, Stian Soiland-Reyes, Simone Leo, Marjan Meurisse, and Henning Hermjakob. By-covid d4. 3 provenance model for infectious diseases. 2024b.
- Laura Wratten, Andreas Wilm, and Jonathan Göke. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature methods*, 18(10):1161–1168, 2021.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024.

- Fanghui Xiao, Daqing He, Yu Chi, Wei Jeng, and Christinger Tomer. Challenges and supports for accessing open government datasets: Data guide for better open data access and uses. In *Proceedings of the 2019 conference on human information interaction and retrieval*, pages 313–317, 2019.
- Chunli Xie, Jerry Gao, and Chuanqi Tao. Big data validation case study. In *2017 IEEE third international conference on big data computing service and applications (BigDataService)*, pages 281–286. IEEE, 2017.
- Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*, 2023.
- L Xue. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.
- Wenzhe Yang, Sheng Wang, Shixun Huang, Yuyang Liao, Yuan Sun, Juliana Freire, and Zhiyong Peng. A unified approach for multi-granularity search over spatial datasets. *arXiv preprint arXiv:2412.04805*, 2024.
- Qinyuan Ye, Maxamed Axmed, Reid Pryzant, and Fereshte Khani. Prompt engineering a prompt engineer. *arXiv preprint arXiv:2311.05661*, 2023.
- JungWon Yoon and EunKyung Chung. An investigation on graphical abstracts use in scholarly articles. *International Journal of Information Management*, 37(1):1371–1379, 2017.
- JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. Why johnny can’t prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2023.
- Jun Zengy, Xiang Wang, Jiahao Liu, Yinfang Chen, Zhenkai Liang, Tat-Seng Chua, and Zheng Leong Chua. Shadewatcher: Recommendation-guided cyber threat analysis using system audit records. In *2022 IEEE symposium on security and privacy (SP)*, pages 489–506. IEEE, 2022.
- Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. Large language models for human-robot interaction: A review. *Biomimetic Intelligence and Robotics*, page 100131, 2023a.
- Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Yan Gao, Yao Hu, and Enhong Chen. Notellm-2: Multimodal large representation models for recommendation. *arXiv preprint arXiv:2405.16789*, 2024.

- Qian Zhang, Yang Cao, Qiwen Wang, Duc Vu, Priyaa Thavasimani, Timothy McPhillips, Paolo Missier, Peter Slaughter, Christopher Jones, Matthew B Jones, et al. Revealing the detailed lineage of script outputs using hybrid provenance. *International Journal of Digital Curation*, 12(2):390–408, 2017.
- Shichao Zhang, Chengqi Zhang, and Qiang Yang. Data preparation for data mining. *Applied artificial intelligence*, 17(5-6):375–381, 2003.
- Shuo Zhang and Krisztian Balog. Ad hoc table retrieval using semantic similarity. In *Proceedings of the 2018 world wide web conference*, pages 1553–1562, 2018.
- Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. *Advances in neural information processing systems*, 36: 76558–76618, 2023b.
- Zitong Zhang and Yaseen Ashraf. A content-based dataset recommendation system for biomedical datasets. In *2023 6th International Conference on Information and Computer Technologies (ICICT)*, pages 198–202. IEEE, 2023.
- Ziyao Zhang, Chong Wang, Yanlin Wang, Ensheng Shi, Yuchi Ma, Wanjun Zhong, Jiachi Chen, Mingzhi Mao, and Zibin Zheng. Llm hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *Proceedings of the ACM on Software Engineering*, 2(ISSTA): 481–503, 2025.
- Jun Zhao, Jose Manuel Gomez-Perez, Khalid Belhajjame, Graham Klyne, Esteban Garcia-Cuesta, Aleix Garrido, Kristina Hettne, Marco Roos, David De Roure, and Carole Goble. Why workflows break—understanding and combating decay in taverna workflows. In *2012 IEEE 8th International Conference on e-Science*, pages 1–9. IEEE, 2012.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022a.
- Huixue Zhou, Robin Austin, Sheng-Chieh Lu, Greg Marc Silverman, Yuqi Zhou, Halil Kilicoglu, Hua Xu, and Rui Zhang. Complementary and integrative health information in the literature: its lexicon and named entity recognition. *Journal of the American Medical Informatics Association*, 31(2):426–434, 2024.

- Wenchao Zhou, Suyog Mapara, Yiqing Ren, Yang Li, Andreas Haeberlen, Zachary Ives, Boon Thau Loo, and Micah Sherr. Distributed time-aware provenance. *Proceedings of the VLDB Endowment*, 6(2):49–60, 2012.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022b.
- Chaim Zins. Conceptual approaches for defining data, information, and knowledge. *Journal of the American society for information science and technology*, 58(4):479–493, 2007.
- Moshe M. Zloof. Query-by-example: A data base language. *IBM systems Journal*, 16(4):324–343, 1977.
- Anneke Zuiderwijk, Marijn Janssen, Sunil Choenni, Ronald Meijer, and Roexsana Sheikh Al-ibaks. Socio-technical impediments of open data. *Electronic Journal of e-Government*, 10(2):pp156–172, 2012.
- Marius Zumwald, Benedikt Knüsel, Christoph Baumberger, Gertrude Hirsch Hadorn, David N Bresch, and Reto Knutti. Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5):e654, 2020.
- Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B Cohen. Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics*, 8(5):358–375, 2007.