

Syracuse University

SURFACE at Syracuse University

Dissertations - ALL

SURFACE at Syracuse University

8-23-2024

Examining Dataset FAIR Compliance in the Research Data Management Lifecycle

Denise L. Devine
Syracuse University

Follow this and additional works at: <https://surface.syr.edu/etd>

Recommended Citation

Devine, Denise L., "Examining Dataset FAIR Compliance in the Research Data Management Lifecycle" (2024). *Dissertations - ALL*. 2009.
<https://surface.syr.edu/etd/2009>

This Dissertation is brought to you for free and open access by the SURFACE at Syracuse University at SURFACE at Syracuse University. It has been accepted for inclusion in Dissertations - ALL by an authorized administrator of SURFACE at Syracuse University. For more information, please contact surface@syr.edu.

Abstract

This thesis investigates the relationship between data curation and FAIR (Findable, Accessible, Interoperable, and Reusable) compliance in research data management, with a focus on the role of metadata. Through case investigations on Data.gov and Google Dataset Search platforms, the research assesses their efficacy in dataset discovery and the impact of metadata. The thesis examines how data curation influences FAIR compliance throughout the research data lifecycle and explores metadata's role in FAIR Principles compliance for both curated traditional research datasets and open datasets. Findings reveal a disparity in FAIR compliance between different dataset types and platforms, with open datasets, particularly those on Data.gov, demonstrating higher compliance due to standardized metadata and formats. In contrast, datasets found through Google Dataset Search exhibit lower compliance levels. While metadata quality generally improves FAIR compliance across repositories, it does not resolve all related issues. The thesis highlights the limitations of heuristic-based approaches in data curation, identifying vulnerabilities such as human error and lack of robust control mechanisms. Results underscore the need for strong data policies to ensure consistent, high-quality research data management practices throughout the data lifecycle.

Keywords: metadata, data curation, FAIR principles, research data management.

Examining Dataset FAIR Compliance in the Research Data Management Lifecycle

by

Denise L. Devine

B.B.A., University of Memphis, 1993

M.S., Syracuse University, 2019

Thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies in Information Management

Syracuse University

August 2024

© Copyright by Denise L. Devine 2024

All Rights Reserved

Dedication

This thesis is dedicated to family and friends whose unwavering support and encouragement I have been blessed with. To my husband Bill, who has been my bedrock of strength and inspiration: your belief in my capabilities and your endless love have been the guiding lights through this journey. To my children, Tiffany, Nicole, and Dustin: your encouragement has made this thesis possible. Plus, I also think you enjoyed the payback from all the times I harassed you during your educational journey. All three of you continually asked how the paper was going and cajoled me into writing—whether I wanted to or not. Face it, you guilted me, yet you were my cheerleaders.

Thank you to my sisters and brother, who cheered me along the way, and provided a sounding board when things were not going well.

I am grateful for Saxon, my big fluffy ball of goof who was always there under my desk, giving me his undivided attention regardless of the time of day or night. His presence is very much missed.

Finally, I cannot forget to thank my friends. In my moments of doubt and stress, your encouraging words and acts of kindness were the beacons that led me back to my path.

Thank you everyone for being my pillars of support and for joining me in celebrating this significant milestone.

Acknowledgments

There are so many people that have made this thesis possible. I would thank Dr. Stephen Wallace for being my advisor in this endeavor. His guidance was so appreciated and needed during this process. So much appreciation and thanks go to Dr. John Jordan; his leadership in this degree program has been outstanding. I must thank my daughter Tiffany who has come to my rescue in editing this thesis. Of course, so many thanks go to Cohort 1; our merry little band of future academics has been fantastic. The friendships we have developed and the mentorship I have received will carry me forward for many years to come.

Table of Contents

List of Tables	x
List of Figures	xiii
CHAPTER 1: INTRODUCTION	1
Background and Context of the Problem.....	1
Research Data Management	1
Open Science	3
Open Data	6
Defining Data.....	8
Metadata.....	9
Frameworks	11
Data Curation	12
Economic Costs of Data	14
Data Policy.....	15
Motivation.....	17
Research Problem Statement	24
Research Objectives and Research Questions	25
RQ1: What role does data curation play in the FAIR compliance of traditional research datasets and how does this role differ in the context of open datasets?.....	25
RQ2: What is the role of metadata in enhancing the compliance of FAIR Principles in curated research datasets versus open datasets?	26
Summary	26
CHAPTER 2: LITERATURE REVIEW	28

Introduction.....	28
Purpose of the Literature Review	28
Scope and Organization	32
Introduction to Research Data Management	33
Definition and Importance of RDM.....	34
Challenges in RDM.....	37
Data Management Lifecycle Models	38
Open Science and Open Data	41
Open Data Initiatives and Policies	42
Benefits and Challenges of Open Data in Research	44
Challenges of Open Data	44
Benefits of Open Data.....	46
Perspectives on What Constitutes Research Data	49
Specific Characteristics of Research Data	52
Open Datasets	54
Traditional Datasets	57
Frameworks for Research Data Management.....	59
FAIR Principles	61
CARE Principles for Indigenous Data Governance.....	62
TRUST Principles for Digital Repositories	65
Data Curation Definition and Importance	67
Key Aspects of Data Curation	67
The Role of Metadata	70

Importance of Metadata for FAIR Compliance	70
Importance of Metadata in Data Curation	71
Importance of Standardized Metadata	72
Differences in Metadata Requirements for Traditional vs. Open Datasets	73
Economic Considerations and Costs in RDM	74
Economic Benefits of Open Data and Open Science.....	76
Data Policies and Their Impact on Research	78
Search and Relevancy	79
Summary	80
CHAPTER 3: METHODOLOGY	82
Research Design	82
Data.Gov	84
Google Dataset Search.....	86
Case Study Setup	87
Phase 1: Evaluating Dataset Findability	87
Phase 2: FAIR Compliance Assessment.....	87
Phase 3: Data Curation Analysis.....	90
Title Search	97
Keyword Search.....	98
Subject Search.....	99
Case Study Recap	100
Validation	100
Documentation.....	102

Summary	104
CHAPTER 4: RESULTS	105
Findings Overview	105
Case Investigation 1 – Title Search	111
Case Investigation 1 – Title Search - Validation	119
Case Investigation 2 – Keyword Search	123
Case Investigation 2 – Keyword Search – Validation	130
Case Investigation 3 – Subject Search	135
Case Investigation 3 – Subject Search – Validation	142
Summary	145
CHAPTER 5: DISCUSSION, RECOMMENDATIONS, and CONCLUSION	162
Discussion of the Findings	162
ISO Policy	166
ISO 25012 Inherent Characteristics and Alignments	168
Application of Policy	174
ISO 25012 Policy Implementation Recommendations	176
Open Research FAIR Compliance Recommendation	177
Dataset Categorization Recommendation	178
Semantic Keyword Search Recommendation	179
Conclusion	180
References	182
VITA	206

List of Tables

Table 1 – RDM Lifecycle Models	39
Table 2 - TRUST Principles for Digital Repositories	66
Table 3 - Dataset FAIR Compliance Evaluation	88
Table 4 – FAIR Compliance Metadata Elements	89
Table 5 – Data Curation Evaluation	91
Table 6 – Data Curation Metadata Fields	93
Table 7 – Metadata FAIR Principles and Data Curation Overlapped Elements.....	96
Table 8 – Validation Process Descriptions	101
Table 9 – Evaluating Dataset Findability Case Investigations Results Overview	110
Table 10 – Case Investigation 1 – Title Search Results.....	111
Table 11 - Case Investigation 1 – Title Search - FAIR Compliance Assessment	114
Table 12 – Case Investigation 1 – Title Search – Data Curation Analysis.....	116
Table 13 – Case Investigation 2 – Keyword Search Results	124
Table 14 - Case Investigation 2 – Keyword Search FAIR Compliance Assessment	126
Table 15 – Case Investigation 2 – Keyword Search – Data Curation Analysis	128
Table 16 – Case Investigation 3 – Subject Search Results	135
Table 17 – Case Investigation 3 – Subject Search – FAIR Compliance Assessment.....	137
Table 18 – Case Investigation 3 - Subject Search – Data Curation Analysis	139
Table 19 – FAIR Compliance Assessment – Data.gov.....	146
Table 20 – FAIR Compliance Assessment – Google Dataset Search	148
Table 21 – Data Curation – Data.gov	151
Table 22 – Data Curation – Google Dataset Search	154

Table 23 – Data.gov – Data Curation Breakdown.....	158
Table 24 – Google Dataset Search – Data Curation Breakdown.....	159
Table 25 - FAIR Principles overlap with data curation metadata.....	164
Table 26 – ISO Based Policy Examples	166
Table 27 - Inherent characteristics of ISO/IEC 25012.....	168
Table 28 – ISO/IEC 25012 and FAIR Principles Alignment	169
Table 29 – ISO/IEC 25012 and Data Curation Guidelines Alignment	170
Table 30 – Review of Issues	173

List of Figures

Figure 1 – Open Science and its Individual Concepts	5
Figure 2 – Motivation Data Example.....	21
Figure 3 –Scopus RDM Publication Trend by Year	29
Figure 4 – Scopus FAIR Principles Publication Trend by Year	30
Figure 5 – Scopus Data Curation Publication Trend by Year	31
Figure 6 – Scopus Data Curation and FAIR Principles Publication Trend by Year.....	31
Figure 7 – FAIR Compliance Analysis - Metadata Collection Datasheet	95
Figure 8 – Data Curation Analysis - Metadata Collection Datasheet Analysis	96
Figure 9 – Evaluating Dataset Findability - Metadata Collection Datasheet.....	102
Figure 10 – FAIR Compliance Assessment – Metadata Collection Datasheet	103
Figure 11 – Data Curation Analysis - Metadata Collection Datasheet	103
Figure 12 – Case Investigation 1 – Validation – Title Search – Data.gov.....	120
Figure 13– Case Investigation 1 – Validation - Title Search - Data.gov	121
Figure 14 – Case Investigation 1 – Title Search – Validation - Google Dataset Search	122
Figure 15 – Metadata discrepancies – Case Investigation 1- Validation.....	123
Figure 16 – Case Investigation 2 – Keyword Search – Validation - Data.gov	131
Figure 17 – Case Investigation 2 – Keyword Search – Validation – Landing - Data.gov	132
Figure 18 – Case Investigation 2 – Keyword Search – Validation - Google Dataset Search.....	133
Figure 19 – Case Investigation 2 – Keyword Search – Google Dataset Search – Alternative Site	134
Figure 20 – Case Investigation 3 – Subject Search – Validation – Data.gov	142
Figure 21 – Case Investigation 3 – Subject Search – Validation - Google Dataset Search.....	143

Figure 22 – AmeriGEOSS site – Case Investigation 3 – Subject Search -Validation.....	144
Figure 23 – FAIR Compliance Criteria	162
Figure 24 – Data Curation Guidelines	163

CHAPTER 1

INTRODUCTION

Background and Context of the Problem

This thesis embarks on an examination of Research Data Management (RDM), a key in the evolving context of open science and open data. Chapter 1 lays the groundwork by defining key concepts and components essential to understanding the lifecycle of research data, from its fundamental definitions to practical aspects like data curation, the economic costs associated with managing data, and the policies that govern data usage. A special focus is placed on metadata and the FAIR framework that support effective data management, underscoring their roles in enhancing data accessibility and reusability.

The motivation behind this study is to address gaps in current RDM practices, which are detailed in the problem statement. This chapter further conveys the research objectives and questions aimed at exploring how strategic data curation can influence the compliance of FAIR (Findability, Accessibility, Interoperability, Reusability) Principles in datasets. A summary concludes the chapter, setting the stage for a deeper investigation into these aspects of data management within the broader scope of scientific research.

Research Data Management

RDM encompasses the organization, storage, preservation, and sharing of data collected and used in a research project. It involves a range of activities, from the planning stage of data collection through to the long-term preservation and dissemination of data. Effective RDM ensures that data is accurate, complete, reliable, and available for future research, thereby maximizing its utility. The importance of RDM has grown with the increase in data-intensive research across various disciplines, making it a critical component of the RDM lifecycle

(Andrikopoulou et al., 2021; Pinfield et al., 2014). To effectively manage this data, structured approaches are essential. According to Ball, data management lifecycle models provide a framework for the various operations that need to be performed on a data record throughout its life (Ball, 2012).

Whyte and Tedds define RDM as a comprehensive term encompassing activities related to the creation, organization, structuring, and naming of data as well as their backup, storage, preservation, and sharing. RDM also includes all actions necessary to ensure data security. Its primary goals are to guarantee the reliable verification of research results and to enable new and innovative research to build on existing data (Schöpfel et al., 2018; Whyte & Tedds, 2011).

RDM involves the practical, day-to-day tasks of handling and processing data within the guidelines established by data stewardship. Data stewardship is a comprehensive approach driven by technology and systematic data management practices. It supports researchers in working collaboratively, sharing ideas, disseminating findings, and reusing results, all while upholding the core values that knowledge should be reusable, modifiable, and redistributable (Arend et al., 2022; Mons, 2018). The National Research Council defines data stewardship to encompass all activities that preserve and improve the information content, accessibility, and usability of data and metadata (G. Peng et al., 2015). RDM and data stewardship are closely interrelated concepts, which when considered together, ensure data is effectively governed, managed, and utilized. They further ensure it creates a robust system for managing data assets, enhancing their value and reliability across the organization (Eaker, 2016).

According to Arend et al. data stewardship involves (i) managing and monitoring the quality of research data as valuable assets and (ii) ensuring the accessibility of high-quality data for the relevant community. Data stewardship is regarded as a component of data management

within the framework of data governance, aiming to manage, curate, and provide data based on user needs (Arend et al., 2022; G. Peng et al., 2015).

Similarly, the Data Driven Life Sciences, a public partnership based on a network of experts and policymakers in the Netherlands, emphasizes that data stewardship includes the proper collection, annotation, and archival of research data as well as its long-term care. Data stewardship is designed to enable researchers to find data and reuse it in downstream studies (Dutch Techcentre for Life Sciences, 2024).

Open science can be an important part of data stewardship because it promotes practices that ensure the transparency, accessibility, and reusability of research data. By enabling more accessible and reproducible research practices, open science aims to enhance the integrity and reliability of scientific findings, which addresses issues related to data silos and fragmented research methods (Stall et al., 2019). This approach enhances the societal impact of research by making scientific discoveries more readily available to a wider audience (National Academies of Sciences, Engineering, and Medicine, 2018).

Open Science

In an era where transparency and collaboration are paramount, open science has emerged as a framework for modern scientific inquiry. This framework seeks to make all aspects of research more accessible, transparent, and collaborative. Open science promotes the idea that scientific knowledge should be freely available to anyone, removing barriers to access and participation. It promotes the unrestricted sharing of research outputs, methodologies, and data to foster greater scientific collaboration and to accelerate the pace of innovation (Fecher & Friesike, 2014; Vicente-Saez & Martinez-Fuentes, 2018). Open science is the movement to make scientific research, data, and dissemination accessible at all levels of an inquiring society. It

represents a change in the way research is conducted, recorded, and disseminated. Open science, therefore, is a paradigm shift in the *modus operandi* of research and science. It impacts the entire scientific process (Ayrís & Ignat, 2018; Suber, 2012).

Open science advocates for the sharing of knowledge, data, and methodologies to facilitate collaboration and accelerate innovation. It encompasses various practices, including open-access publishing, open data, and open methodology, and aims to enhance the reproducibility and integrity of scientific research (Ayrís & Ignat, 2018; Spicer, 2018). Over the past 20 years, several open data movements have been created worldwide, driven by policies such as the Public Sector Information Directive in Europe, the U.S. Open Data Initiative, the Open Government Partnership, and the G8 Open Data Charter. These initiatives promote transparency, accountability, and reuse of data, enabling public participation in decision-making and policy evaluation (Attard et al., 2015; Davies & Perini, 2016).

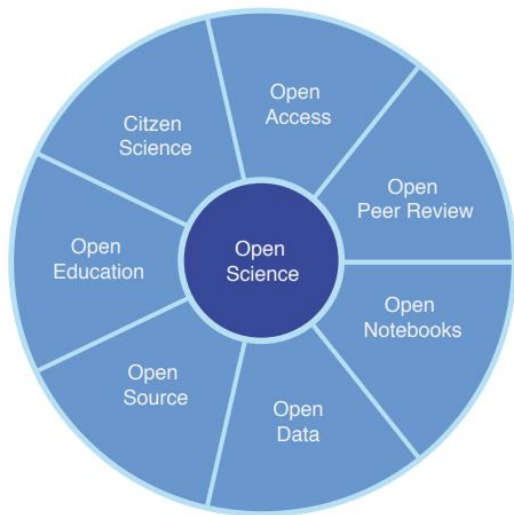
The push toward open science has been significantly influenced by funding bodies like the National Institutes of Health (NIH), the National Science Foundation (NSF) and the European Commission, among others. In line with this movement, universities and research institutions are implementing policies and developing the necessary infrastructure to promote data transparency, advocating for unrestricted access to scientific data. Proponents of open science champion the free access and use of scholarly research outputs, including publications, data, methodologies, and software, thus emphasizing their critical role in enhancing research management (Arpin & Kambesis, 2022; Tsueng et al., 2023).

The following diagram from Rachael Ann Spicer's work "Fit for Purpose? A Meta Scientific Analysis of Metabolomics Data in Public Repositories" provides an overview of the various individual segments that make up the open science concept. These individual concepts,

such as open access, open peer review, open notebooks, citizen science, open education, open source, and open data are shown in Figure 1 below.

Figure 1

Open Science and its Individual Concepts



Note: Adapted from “*Fit for purpose? A meta scientific analysis of metabolomics data in public repositories*” by R. A. Spicer, 2018, University of Cambridge, p. 11.

(<https://doi.org/10.17863/CAM.34945>). In the public domain.

Spicer’s open science model illustrates how open access and open data are not necessarily the same. In the book *Open Access*, author P. Suber notes that open access differs from what he considers to be sub-topics of open data—open educational resources, open government, free and open-source software, or open science (which combines open access texts, open data, and open-source software, and provides this openness at every stage of a research project) (Suber, 2012). Open access is typically the outcome of research that is published and readily available, while open data is the data used in the research that generated the published outcome (Arzberger et al., 2004; Tennant et al., 2016).

Former U.S. President Obama's Open Data Initiative was kickstarted by his Open Government Memorandum in March 2009. The memo preceded the Open Data Policy – an executive order – which was signed in 2013. Its goal was to make information resources accessible, discoverable, and usable by the public to fuel entrepreneurship, innovation, and scientific discovery (Borgesius et al., 2015). The executive order was also designed to commit department heads to publish data and documents within a certain timeframe. This data must be in accordance with guidelines “ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by the agency” (Coglianese, 2009). Also in 2013, the United States Office of Science and Technology mandated that the results and research products of federally funded research be publicly accessible (Arpin & Kambesis, 2022).

Open Data

Open data represents a crucial element of the open science ecosystem, and covers a broad spectrum including open administrative data, open government research data, and open science research data (Spicer, 2018). Beyond its integral role in transparency and collaboration, open data also offers practical benefits, including cost savings. It eliminates redundant efforts and optimizes resource allocation by making data readily available for reuse and analysis, thereby helping organizations, especially in the U.S. government and public sectors, forego the expenses associated with data collection and analysis. This, in turn, facilitates more informed decision-making and highlights opportunities for cost reduction and efficiency enhancement (National Institutes of Health, 2023).

For open data to be truly effective, it must adhere to specific standards that guarantee it accessibility, reliability, accuracy, and security. These standards encompass the provision of data

in a timely and accessible manner; the use of clear methodologies; ensuring data anonymization for privacy; maintenance of data accuracy and completeness; ensuring reliability; implementation of robust security measures; and ensuring long-term preservation (2.15 Open Government Data Act—2018, 2018; Bertagnolli et al., 2017; Van Noorden, 2013).

The interest in open data is escalating across various domains, including genomics, bioinformatics, astronomy, climate science, environmental studies, social sciences, and economics. Open data necessitates that data not only be freely accessible but also extant in a format that facilitates modification and adaptation. It should be shared under conditions that allow for its reuse and redistribution, including integration with other datasets. This ensures that open data is not just available, but also broadly usable by anyone interested (Open Knowledge, 2022; Thorsby et al., 2017)..

The true value of open data lies in its shareability and the ease with which it can be accessed by others, rather than its mere quantity. This approach meets technical, economic, and legal benchmarks by ensuring data is freely accessible online in a user-friendly format that supports further utilization (Chignard, 2013). Access to and sharing of data are pivotal for scientific progress. Consequently, the entire scientific community, from top-tier institutions to individual researchers, should embrace changes that enhance the collective pool of knowledge and understanding produced by research endeavors (Arzberger et al., 2004).

A gap exists between the current practices and the ideals of open data. It can be attributed to several factors: (i) uncertainty among researchers about their legal rights to self-archive data, (ii) concerns that self-archiving might jeopardize the acceptance of their work for publication, (iii) the perception that self-archiving demands significant effort, and (iv) the costs associated

with data acquisition. Regardless of whether data is openly available or restricted, it is essential to recognize its value as an asset (Patel, 2016; Tennant et al., 2016).

Defining Data

Data is a fundamental element in research and decision-making, representing raw facts and figures collected through various methods. It consists of unprocessed observations and measurements that can be quantified and analyzed, taking numerous forms such as numbers, text, images, or sounds. These are gathered from diverse sources including experiments, surveys, and digital traces (Borgman, 2021). The significance of data lies not merely in its raw form, but in its potential to be transformed into meaningful information and knowledge through analysis and interpretation (Chignard, 2013).

Data can be categorized into several types, each with unique characteristics and applications. Quantitative data refers to numerical information that can be measured and statistically analyzed, such as temperature readings or survey responses. It includes, counts, measurements, and other data expressed in numbers, like laboratory measurements, and census data (Caulley, 2007; Creswell & Poth, 2024). In contrast qualitative data encompasses non-numerical information, like interview transcripts or video recordings, which require more interpretive analysis to uncover patterns and insights. Qualitative data can consist of textual data (e.g., interview transcripts, open-ended survey responses), visual data (e.g., photographs, videos), and audio data (e.g., recordings of interviews, focus groups) (Creswell & Poth, 2024).

Research data, also known as scientific or scholarly data, is a specific subset of data crucial to scientific work. It can be defined as information in documents or digital form collected, observed, created, or produced during scientific research. Research data is essential for generating insights, ensuring reproducibility, and validating findings (Caulley, 2007; Gregory et

al., 2018). It comes in various forms and from multiple sources, each suited to different types of inquiries and disciplines. The most common forms of research data are quantitative, qualitative, and mixed methods (Creswell & Poth, 2024).

The nature and context of research data can vary across disciplines. In the biological sciences, chemistry, and crystallography, research data is often derived from instruments and is highly structured (Langer & Bilz, 2019). In contrast, the digital humanities and social sciences frequently use large quantities of textual data that may need to be transcribed and encoded for further processing (Allen & Hartland, 2018). Regardless of the field, research data must be stored, organized, documented, preserved (or discarded), and made discoverable and (re)usable (Owan & Bassey, 2019; Qin et al., 2017).

In the scientific community, research data is the backbone of scientific inquiry. Science, social science, and the humanities are increasingly data-intensive, highly collaborative, and computational at a large scale (Borgman, 2012; Qin et al., 2017). As the volume and complexity of data continues to grow, the implementation of robust data practices becomes increasingly important for maintaining the quality, integrity, and utility of research data (Mons, 2018; Wilkinson et al., 2016).

Metadata

Metadata, often described as “data about data” plays a role in the organization, discovery, and utilization of information in the digital age (Arpin & Kambesis, 2022; Burton & Treloar, 2009). While this simple definition captures the principle of metadata, its importance and complexity extend far beyond this basic description. Metadata serves as the invisible infrastructure that underpins our ability to navigate, understand, and utilize the vast amounts of research data that characterizes modern research and information management. Its role in

improving discoverability, facilitating interoperability, and ensuring the long-term usability of research data makes it an important component of the RDM ecosystem (Whyte & Tedds, 2011).

Jenn Riley, in her book *Understanding Metadata*, offers a more comprehensive definition, describing metadata as the information we create, store, and share to describe things. This definition emphasizes metadata's active facilitation role in interactions with digital objects and how it enables users to access the knowledge they see. Riley further explains that metadata is fundamental to the functionality of content-holding systems, allowing users to find items of interest, record essential information about them, and share that information with others (Riley, 2017).

The significance of metadata extends across a broad range of applications, serving as a keystone component that connects contents and services, thereby enhancing their visibility and accessibility to users (Weibel & Koch, 2000). In the area of digital resources, where the sheer volume of information can be overwhelming, metadata becomes central to the objective of finding and retrieving relevant material. It provides a structured framework for describing digital objects, making them discoverable and manageable within complex information systems (Weagley et al., 2010).

Metadata standards have evolved to provide consistent ways for researchers to describe their projects and datasets. These standards not only ensure uniformity in data description, but also make datasets machine-readable, enabling them to be indexed and searched efficiently (Eaker, 2016). There are many metadata standards. The Dublin Core schema, for example, emerged as a widely used metadata standard for electronic resources. It is designed to facilitate high-level searching of textual documents across various disciplines, databases, and schemas (Hunter & Iannella, 1998; Weagley et al., 2010).

In the context of RDM, metadata plays a key role in indexing and discovering research information. Persistent identifiers (PIDs) and metadata schemas are crucial components of effective RDM practices, enabling the long-term accessibility and usability of research data (Hauschke et al., 2021). The development of metadata standards has also paved the way for implementing the concept of linked data, further enhancing the interoperability and discoverability of digital resources across different domains (Eaker, 2016).

However, the application of metadata in research contexts requires a balance. Quimbert et al. discuss the need to provide sufficient general information for broad searching and access to research data, while also including specific information that caters to the unique requirements of different research communities. The concept of linked data has emerged as a promising approach to improve this equilibrium, offering enhanced interoperability between different domains and improving the metadata landscape (Quimbert et al., 2020).

Frameworks

Effective RDM is one of the keys to ensuring the integrity, accessibility, and long-term usability of scientific data (Borgman, 2013; Whyte & Tedds, 2011). To address the diverse needs and challenges in this domain, several frameworks have been developed, each with its unique focus and guiding principles. These frameworks include FAIR, CARE, and TRUST principles. This thesis will focus on FAIR Principles. However, it will also introduce the characteristics the CARE and TRUST principles contribute to the RDM ecosystem.

The FAIR Principles, established by Wilkinson et al., emphasize that data should be Findable, Accessible, Interoperable, and Reusable. These principles aim to enhance the utility and impact of data by ensuring it is easily discoverable, accessible under well-defined conditions, compatible with other datasets, and suitable for reuse in future research. The FAIR

framework advocates for the use of standardized metadata, persistent identifiers, and open data formats to achieve these goals, thereby fostering greater transparency and collaboration within the scientific community (Mons, 2018; Wilkinson et al., 2016).

Complementing the FAIR Principles, the CARE Principles were developed by the Global Indigenous Data Alliance (GIDA) to address the specific needs and rights of Indigenous communities in data governance. CARE stands for Collective Benefit, Authority to Control, Responsibility, and Ethics. This framework highlights the importance of recognizing Indigenous sovereignty over their data, ensuring that data practices respect Indigenous cultures and knowledge systems, and promotes the ethical use of data to benefit Indigenous peoples (Carroll et al., 2020).

Finally, the TRUST Principles provide guidelines for creating and maintaining trustworthy digital repositories. TRUST stands for Transparency, Responsibility, User focus, Sustainability, and Technology. These principles are designed to ensure that digital repositories maintain high standards of transparency regarding their data policies and practices; take responsibility for the long-term stewardship of data; prioritize the needs and expectations of their users; operate sustainably over time; and employ robust technologies to safeguard data integrity (Lin et al., 2020).

By applying any or all of the principles highlighted above, FAIR, CARE, and TRUST in the RDM process, researchers, institutions, and data repositories can ensure that research data is findable, accessible, interoperable, and reusable, while adhering to ethical and governance standards, promoting sustainability, and respecting the rights and interests of Indigenous peoples and other stakeholders (Carroll et al., 2020; Lin et al., 2020; Mons, 2018; Wilkinson et al., 2016).

Data Curation

Data curation encompasses the comprehensive management of data from its inception, ensuring it remains reliable, accessible, and suitable for its intended uses. This process involves not only the preservation of data but also its active maintenance and enhancement, including continuous updates to keep dynamic data fit for purpose. Proper data curation is crucial as it directly influences the availability and utility of data for the research community and adheres to high standards set by frameworks like the FAIR Principles. Without proper curation, data can become obsolete, unreliable, or inaccessible, undermining the research process and principles of open science (Gonzalez & Peres-Neto, 2015; Lord et al., 2004; Sheridan et al., 2021).

Key components of data curation include:

Data Organization: Structuring and categorizing data in a logical, coherent manner to facilitate navigation and retrieval. This involves classifying data into hierarchies, tagging, and using standard taxonomies (Broman & Woo, 2018; Gilliland, 2008).

Data Quality: Verifying data accuracy, consistency, and completeness. High-quality data is essential for reliable research and decision-making. Effective curation practices identify and rectify errors, inconsistencies, or gaps in datasets (Batini et al., 2009).

Metadata Creation: Providing detailed and accurate metadata, describing the data's origin, purpose, creation time, creator, location, and format. Metadata enhances data discoverability and facilitates understanding of the data context and constraints (Duval et al., 2002; Palmer, 2009).

Archiving: Storing data securely for long-term retention, ensuring it remains available for future research, and safeguarding against data loss due to technological obsolescence or other risks (Beagrie, 2006).

Preservation: Maintaining data over time, ensuring its usability and accessibility despite changes in technologies and formats. This includes converting data into less likely obsolete formats and ensuring its physical and digital security (Keene, 2002).

Accessibility: Making data available to its intended users through intuitive, easy-to-navigate interfaces and access systems. Ensuring data availability respects privacy and ethical considerations (Jati et al., 2022).

Effective data curation ensures research data remains a robust and valuable resource for scientific inquiry, balancing technical, practical, and ethical aspects to support ongoing and future research efforts (Beagrie, 2006; Jacobsen et al., 2020; Jati et al., 2022; Wilkinson et al., 2016).

Economic Costs of Data

Implementing frameworks like the FAIR Principles, the CARE Principles, or TRUST Principles incurs costs. These include training people to manage and curate data effectively, developing processes that ensure data quality and compliance with standards, investing in technologies that facilitate data storage and sharing, and maintaining the data infrastructure. Additionally, there are economic costs associated with treating data as an asset. Organizations must invest in data stewardship roles and activities to ensure data's long-term value (Arend et al., 2022). This includes expenditures on data curation, quality control, and infrastructure to manage and protect data assets. These investments, though significant, are necessary to maximize the utility and economic value of data in research (Allen & Hartland, 2018; Arend et al., 2022; R. Peng, 2015).

Management should be persuaded that adopting frameworks like FAIR Principles will yield a long-term return on investment (Directorate-General for Research and Innovation

(European Commission), 2018). Costs associated with these frameworks can be justified by the strong potential for improved data reuse, enhanced collaboration, and accelerated innovation (Harrow & Liener, 2021). When data is managed as an asset, it can drive efficiency and create new opportunities for research and development, leading to economic and societal benefits. Investing in any of the aforementioned principle frameworks can result in more efficient and effective use of data across the research community (Carroll et al., 2021; Jacobsen et al., 2020; Lin et al., 2020; Wilkinson et al., 2016).

In the context of RDM, the economic benefits of open data and open science are multifaceted. Open data initiatives drive innovation by providing researchers with access to a large number of datasets, fostering development of new products, insights, and services that can lead to economic growth (Manyika et al., 2013). By reducing redundancy and increasing efficiency, open data can result in substantial cost savings for both public and private sectors, enhancing the overall efficiency of research processes (Directorate-General for Communications Networks et al., n.d.). Additionally, open science accelerates the pace of research and development by facilitating data sharing and collaboration, which can lead to faster scientific discoveries and technological advancements (Paic, 2021). Ultimately, open data and open science contribute to economic growth and job creation by creating new markets and opportunities within data-driven industries (Tennant et al., 2016).

Data Policy

The adoption of the FAIR Principles, the CARE Principles, and TRUST Principles, combined with data curation practices within the RDM lifecycle, is significantly influenced by policies that promote open science and data sharing. For instance, the NIH and the NSF in the United States, as well as the European Commission, have implemented policies that require

researchers to share their data openly. These policies aim to enhance transparency, reproducibility, and the overall integrity of scientific research by ensuring that data is accessible and reusable (Bertagnolli et al., 2017; National Institutes of Health, 2023).

Policies such as the U.S. Open Data Policy and the European Open Science Cloud initiative underscore the importance of making research data available to the public. They mandate that data generated from publicly funded research should be freely accessible, ensuring that valuable data is not wasted and can be used for secondary analysis (Arpin & Kambesis, 2022). Effective data policies play a role in RDM by establishing standards and practices that promote data sharing, accessibility, and preservation. Here are some examples of data policies and their impact on research:

National Institutes of Health Data Sharing Policy: The NIH has implemented a data-sharing policy that requires researchers to submit a data-sharing plan for grants exceeding \$500,000. This policy aims to enhance data accessibility and facilitate the reuse of data to advance scientific discovery. The impact of this policy includes increased collaboration among researchers, more efficient use of research funds, and accelerated scientific progress by enabling secondary analysis of existing data (National Institutes of Health, 2023).

European Union's General Data Protection Regulation (GDPR): The GDPR establishes stringent guidelines for data protection and privacy, impacting how research involving personal information is managed. While primarily focused on privacy, GDPR also emphasizes the importance of data documentation, transparency, and accountability. Its impact on research includes heightened awareness and better practices regarding data security, ethical data handling, and increased trust among research participants (Radley-Gardner et al., 2016).

Research Councils UK (RCUK) Common Principles on Data Policy: The RCUK has established common principles that mandate researchers to make their data openly available with as few restrictions as possible, while also protecting the legitimate interests of researchers and participants. These principles have led to increased data sharing and collaboration, improved reproducibility of research findings, and greater transparency in the research process (Hodson, 2011).

NSF Data Management Plan Requirements: The NSF requires grant applicants to submit a data management plan (DMP) outlining how data will be managed and shared. This policy encourages researchers to plan for data sharing and preservation from the outset of their projects, leading to better organized and more accessible datasets. The impact includes enhanced data stewardship, increased opportunities for data reuse, and greater scientific collaboration (Foundation, 2017).

These examples of policies collectively enhance RDM by promoting data sharing, ensuring data quality and reproducibility, and enhancing data security and ethical standards. They also increase efficiency and reduce costs. These best practices in RDM are important in advancing scientific research and advancing a collaborative research environment (Foundation, 2017; Hodson, 2011; Manyika et al., 2013; Radley-Gardner et al., 2016).

Motivation

This section explores a citizen-science, marine-based project operated by a non-profit organization, and examines its data collection process. The team's selection of the FAIR Principles as the framework for analysis in this project is particularly relevant. The FAIR Principles provide a comprehensive and widely recognized set of guidelines that address the key challenges faced in research data management. In the context of this non-profit's project, in

which data collection and storage practices were found to be lacking in standardization and structure, the FAIR Principles offer a roadmap for improvement. They provide a structured approach to enhancing data organization, accessibility, and reusability—critical aspects for citizen science projects, in which data sharing and long-term usability are often key objectives. By analyzing the project’s data management practices against the FAIR Principles, the author could effectively identify gaps in the current system and propose targeted improvements to improve the quality, accessibility, and potential impact of the research data collected on this marine-based project.

The examination of the research team’s data highlights the need to adhere to FAIR Principles guidelines and data curation management processes. Specific details about the project were withheld due to confidentiality agreements. Additionally, analysis of the team’s data organization was conducted under a non-disclosure agreement (NDA) signed by the author, who assessed the non-profit’s research data collection and storage methodology in a consultancy capacity. This involved recommending strategies to enhance data organization and identifying needs for RDM during a meeting with the project’s lead researcher.

A team of citizen scientists has collected 10 years’ worth of debris picked up on beaches. As part of its collection efforts, the team tagged a location based on the map name of a particular beach—but did not use GPS coordinates. Then, team members sorted and described the objects they removed from the beach. However, the team did not utilize a formal classification method. Nor did members employ a master list of classifications, such as the one provided by the National Oceanic and Atmospheric Administration (NOAA).

For example, if researchers found a piece of plastic, they described it with minimal characteristics, such as “white plastic found on XYZ beach” on a specific date. The specifics of

members' data were stored in Google Sheets in a series of separate spreadsheets. Specific characteristics could be noted on a Google Sheet and listed by location, or by type of plastic, therefore duplicating the data. An example of this approach was to have a Google Sheet with all debris found on XYZ beach, then a second Google Sheet with the same information, but broken down by plastic characteristics— “blue plastic,” for instance. Team members who wished to review a spreadsheet would copy it, then edit the duplicate (if necessary), and then repeat those changes on the original.

The lead researcher did not want to share data because she was afraid people might steal it and do their own analysis, for their own papers. When the lead researcher did offer her data to other research groups, it was rejected because the research data was not in a normalized format. When asked about developing a data repository of some kind, the lead researcher balked. She indicated a repository was not necessary because the data could not be normalized from Google Sheets without intensive data cleansing. The actual objects physically currently exist in storage, but the metadata that could have been collected at the time of the initial retrieval was lost.

To recap, during the research lifecycle phase of creating data/data describing, these citizen scientists did not use standardized methods such as “categorization” within their domain of beach debris collected. Because the objects were not identified using domain categorization, many characteristics were missing. The lead researcher was concerned that data she had collected and categorized using her own categories could be used in an unauthorized manner, without her consent. In fact, data was recorded within an insecure medium, where any user with access could copy it—and change the copy—without changing the original. Such a user could also inadvertently change the original data because there were no safeguards against such

alterations. And, if a researcher left the program, there were no indications of changes made (if any). Such security breaches can lead to data provenance gaps and enable data corruption.

These data collection issues (within the research lifecycle) could affect the data processing phase because when data records stored in the Google Sheets are gathered, the data lacks clear provenance. The opportunity to gather metadata has been, at the very least, minimized. Without the security of knowing the data record is in its original state, data analysis performed in another research phase is suspect—and possibly unreliable. Overall, because the data is not stored in a formal repository such as a database, it cannot be reused by other scientific groups. A lack of data curation data management is one of several components missing from this open science-based research. Findability, even if limited to internal users, is hindered due to missing metadata and metadata standards. The ability of researchers to control the long-term storage of data is also absent. Managing data means maintaining necessary context information and associated documentation to ensure other researchers (and the original data owners) can use the data when the need arises. Good curation means good science (Rusbridge, 2007). An example of the citizen's data is shown in Figure 2. It has been anonymized to ensure the NDA was not breached.

Figure 2

Motivation Data Example

1	Marine Debris Removal Data Sheet									
2		Location:	My Secret Beach			Date:	20 October 2021			
3										
4		Total whole items =	27	Eaten pieces =	1	Weight (kg) =	1.98 kg			xx indica
5		Total pieces =	22	Other pieces =	21	Weight (lb) =				5.65 kg
6		Total items =	49	Total pieces	22					
7										
8		INDUSTRY/ACTIVITY			TYPES OF MARINE DEBRIS			MATERIAL		WHOLE
10		Fishing			Basket			plastic		
11		Fishing			Buoy			plastic		
12		Fishing			Float - other			plastic		
13		Fishing			Miscellaneous			plastic		
14		Aquaculture			Seaweed ring			plastic		
15		General Ocean Industry			Grate			plastic		
16		Food/Drink			Beverage bottle			plastic		
17		Food/Drink			Cap - Nestle			plastic		
18		Food/Drink			Rice spatula			plastic		
19		Personal Care/Hygiene			Comb - other			plastic		
20		General Living/Household			Umbrella - handle only			plastic		
21		Toys/Games			General toy			plastic		
22		Toys/Games			Toy wheel			plastic		
23		Tobacco/Smoking			Lighter			plastic		
24		Unknown Industry/Activity			Bottle - other			plastic		
25		Unknown Industry/Activity			Bottle - bottom only			plastic		
26		Unknown Industry/Activity			Cap/lid (plastic)			plastic		
27		Fishing			Twine spool			plastic		

Note: Motivation Data Example is a purely visual depiction of how the datasheet was constructed. There is no actual real data presented.

Reviewing the issues identified in the example, the application of the FAIR Principles and data curation practices to this citizen-science marine-based project could significantly improve their project in several ways:

1. Enhanced Findability:

- **Current Issue:** The data is stored in Google Sheets without unique identifiers or standardized metadata, making it difficult to locate specific datasets efficiently.
- **FAIR Solution:** Assign globally unique and persistent identifiers to each dataset and use rich metadata descriptions. This would allow both current and future researchers to easily find the data, enhancing the project's transparency and usability.

2. Improved Accessibility:

- **Current Issue:** Data is duplicated across multiple spreadsheets, and access is controlled informally, leading to potential data loss and inconsistencies.
- **FAIR Solution:** Store data in a standardized format within a secure, accessible repository. This ensures that data is retrievable by authorized users under well-defined conditions, and metadata remains accessible even if the data itself is no longer available.

3. Better Interoperability:

- **Current Issue:** Data is not categorized using standardized methods, resulting in inconsistencies and difficulties in integrating with other datasets or systems.
- **FAIR Solution:** Adopt formal, accessible, shared vocabularies and classification systems, such as those provided by NOAA. This enables data to be easily integrated and used in conjunction with other datasets, facilitating broader scientific inquiries and collaborations.

4. Increased Reusability:

- **Current Issue:** Lack of detailed metadata and a standardized format makes it challenging for other researchers to understand and reuse the data.
- **FAIR Solution:** Ensure data and metadata are richly described with accurate and relevant attributes, including provenance information and clear usage licenses. This makes the data more comprehensible and useful for future research, maximizing its value and impact.

5. Enhanced Data Curation:

- **Current Issue:** Data management practices are informal and insecure, leading to potential data corruption and loss of provenance.

- **Data Curation Solution:** Implement thorough data curation practices, including data selection, preservation, refinement, and archiving. Create detailed metadata, improve data quality, standardize data formats, and establish transparent access policies. This ensures the long-term usability and integrity of the data.

6. Promoting Open Science:

- **Current Issue:** The lead researcher's reluctance to share data due to fear of misuse, combined with the non-normalized format of data, hinders collaboration and broader scientific impact.
- **Open Science Solution:** By adopting FAIR Principles and data curation practices, the project can securely share data, enhancing collaboration and transparency. This aligns with the principles of open science, promoting broader dissemination and use of research data.

7. Addressing Critiques and Challenges:

- **Implementation Challenges:** While implementing FAIR can be resource-intensive, especially for smaller research groups, seeking support from funding bodies and collaborations with larger institutions can mitigate these challenges.
- **Uneven Adoption:** To avoid fragmentation, it's important to promote standardized practices across disciplines through training, resources, and community engagement.

By adopting FAIR Principles and robust data curation practices, the citizen-science project can transform its RDM processes, enhancing the reliability, accessibility, and impact of

its research. Doing so would not only improve the project's outcomes, but also contribute to the broader scientific community's efforts in promoting open, transparent, and reproducible science.

Research Problem Statement

The increasing complexity and volume of data-intensive research across various disciplines have highlighted the critical need for effective RDM. Effective RDM encompasses the organization, storage, preservation, and sharing of data collected and used in research projects, ensuring data accuracy, completeness, and reliability for future research endeavors (Andrikopoulou et al., 2021; Pinfield et al., 2014). Despite its importance, there are significant challenges in ensuring that research data is managed in accordance with the FAIR Principles (Findable, Accessible, Interoperable, Reusable), which are essential for maximizing data utility and enabling new and innovative research (Jacobsen et al., 2020; Wilkinson et al., 2016).

A fundamental issue identified is the inadequate data curation practices, which play a role in ensuring data complies with the FAIR Principles. Data curation involves processing such as categorization, organization, and storage of datasets, which are essential to maintain data integrity and usability. However, the lack of comprehensive curation practices often leads to incomplete metadata and poor data quality, hampering the discoverability and reusability of research data (Schöpfel et al., 2018; Whyte & Tedds, 2011).

The problem is exemplified in various settings, including publicly accessible data repositories and internally stored datasets, such as those found in citizen-science projects. Inconsistent data collection and curation practices in these settings often result in data that lacks proper documentation and standardization, leading to significant gaps in data provenance and usability (Wiggins & Wilbanks, 2019). Additionally, even within established data repositories,

discrepancies in curation policies and practices often exclude datasets that do not conform to specific standards, further complicating data integration and interoperability (Borgman, 2013).

Moreover, the absence of universally accepted metadata standards presents a significant barrier to effective data management. Detailed and standardized metadata are crucial for data discovery, providing essential information that helps researchers locate and access the suitability of datasets for specific research purposes (Palmer, 2009). The lack of such metadata standards hinders interdisciplinary research and the seamless use of datasets across various scientific domains (Jacobsen et al., 2020; Wilkinson et al., 2016).

In summary, this research problem underscores the multifaceted challenges associated with data curation and the application of the FAIR Principles in datasets. It highlights the need for improved curation practices and the application of the FAIR Principles within the RDM environment. It also underscores the need for policies to provide overall guidance in RDM processes. This thesis will examine the roles of the FAIR Principles and data curation, and how applying these concepts to datasets can enhance their utility.

Research Objectives and Research Questions

This thesis's motivation—and oft-appearing research problems—highlights the challenges researchers face using the FAIR Principles as guidelines, as well as those of data curation. These challenges stem from not knowing when a dataset follows the FAIR Principles and the FAIR compliance of the dataset, and if data curation processes were used. This thesis aims to address the following research questions to understand and identify viable solutions for these challenges:

RQ1: What role does data curation play in the FAIR compliance of traditional research datasets and how does this role differ in the context of open datasets?

Expectations: Data curation is expected to significantly enhance the FAIR Principles of traditional research datasets through systematic organization, categorization, and metadata management. In contrast, while data curation remains vital for open datasets, it requires a distinct approach that emphasizes the standardization of formats and metadata. This distinction is necessary to ensure broader accessibility and interoperability, catering to the diverse needs of the open data community (Palmer, 2009; Renear et al., 2010; Renear & Palmer, 2009).

RQ2: What is the role of metadata in enhancing the compliance of FAIR Principles in curated research datasets versus open datasets?

Expectations: This research will explore the premise that metadata is a key element in ensuring compliance of FAIR Principles and usability of both curated traditional research datasets and research open datasets. The expectation is that the role of metadata is more pronounced in open datasets due to a greater necessity for standardization and comprehensive metadata practices. These practices are essential for facilitating data integration and reuse across different platforms and disciplines. In curated traditional research datasets, while metadata is integral, the controlled environment may afford more flexibility in metadata structuring and application (Brickley et al., 2019; Devaraju & Berkovsky, 2018).

Summary

This thesis emphasizes that while the application of the FAIR Principles (Findable, Accessible, Interoperable, Reusable) is important, they are insufficient by themselves. It highlights the importance of effective data curation practices in conjunction with the FAIR Principles to achieve improvements in RDM. The research underscores the important role data curation plays in ensuring that datasets adhere to the FAIR guidelines, thereby improving their utility and accessibility for future research. It highlights the multifaceted challenges associated

with managing research datasets, emphasizing the need for comprehensive and standardized metadata and data management policies.

The analysis includes a case study of a citizen-science marine-based project, which reveals significant gaps in data collection, classification, and storage practices. These gaps underscore the importance of following standardized protocols and utilizing robust data management frameworks to ensure data quality and reusability. By investigating the relationship between data curation and the FAIR Principles, the thesis aims to provide insights into best practices for data management that can be applied across various scientific disciplines.

In addressing these challenges, the thesis outlines key research questions focused on understanding the role of data curation in improving the FAIR compliance of research datasets and the importance of metadata in supporting these principles. The expected outcomes include improved methodologies for data organization, enhanced data sharing practices, and the development of policies that support efficient and effective RDM. This research contributes to the broader understanding of how to manage research data in a way that maximizes its value and utility for scientific advancement.

CHAPTER 2

LITERATURE REVIEW

Introduction

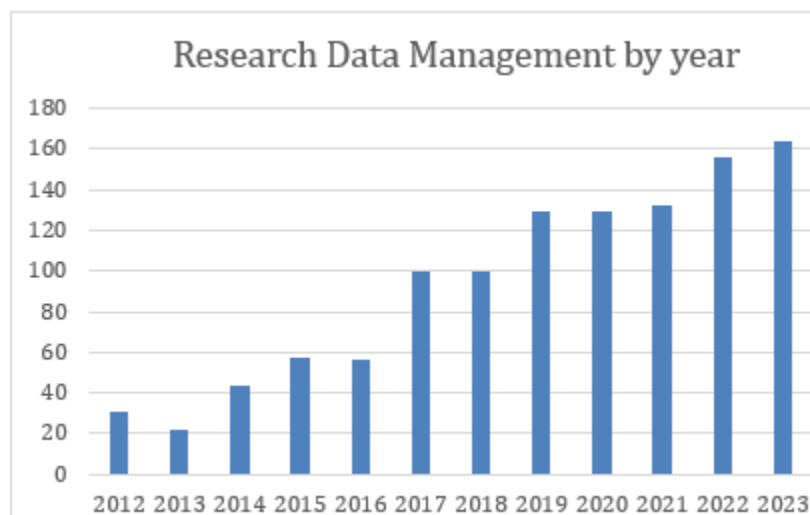
This literature review encapsulates a range of scholarly perspectives on the intersection of open science, open data, FAIR Principles, and data curation, as seen through the lens of RDM. By examining these key areas, this thesis will contribute to the broader discourse of effective data management practices by identifying and improving the overall FAIR compliance of scientific research data. Using the RDM framework, this thesis will also identify issues in data curation and provide suggestions to improve data curation practices.

Purpose of the Literature Review

One of the first steps to understanding how the RDM field is being used is to determine what academic studies have been done on the subject. Entering the keywords “Research Data Management” into a Scopus search engine yields 1,223 documents. The publication dates range from 1945 to 2024. The years from 1945 to 2011 averaged two articles per year. However, starting in 2012, the number of published articles jumped substantially. In 2011, three articles were published; in 2012 that figure jumped to 31—a clear indication that “Research Data Management” was becoming a more important part of the research lifecycle (Andrikopoulou et al., 2021). The following graphic, Figure 3—Scopus RDM Publication Trend by Year, illustrates the upward trend in RDM-related publishing.

Figure 3

Scopus RDM Publication Trend by Year

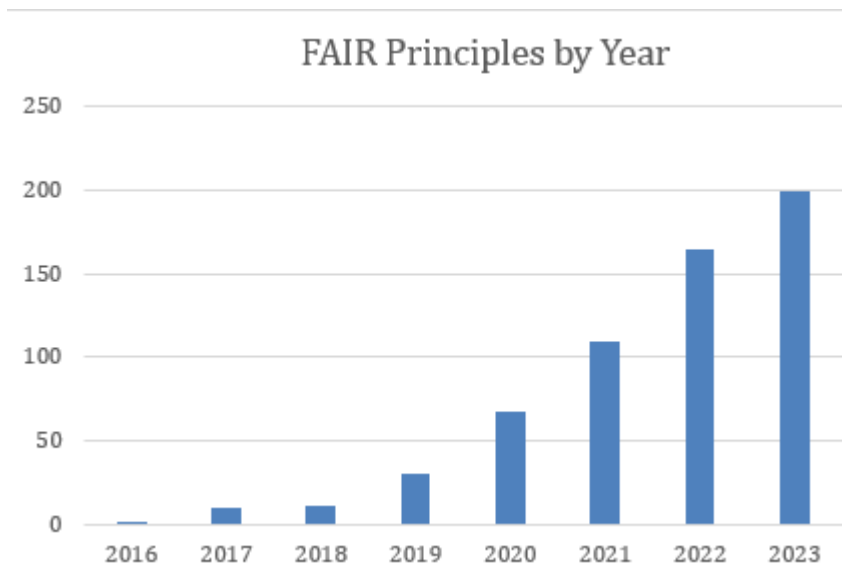


The same exercise can be repeated using “FAIR Principles” as the search parameter.

Those keywords produced 684 documents meeting the search criteria. Further search refinements included searching for “FAIR Principle” and “FAIR.” Published articles on these topics also increased significantly after 2011, as shown by Figure 4, “Scopus FAIR Principles Publication Trend by Year.”

Figure 4

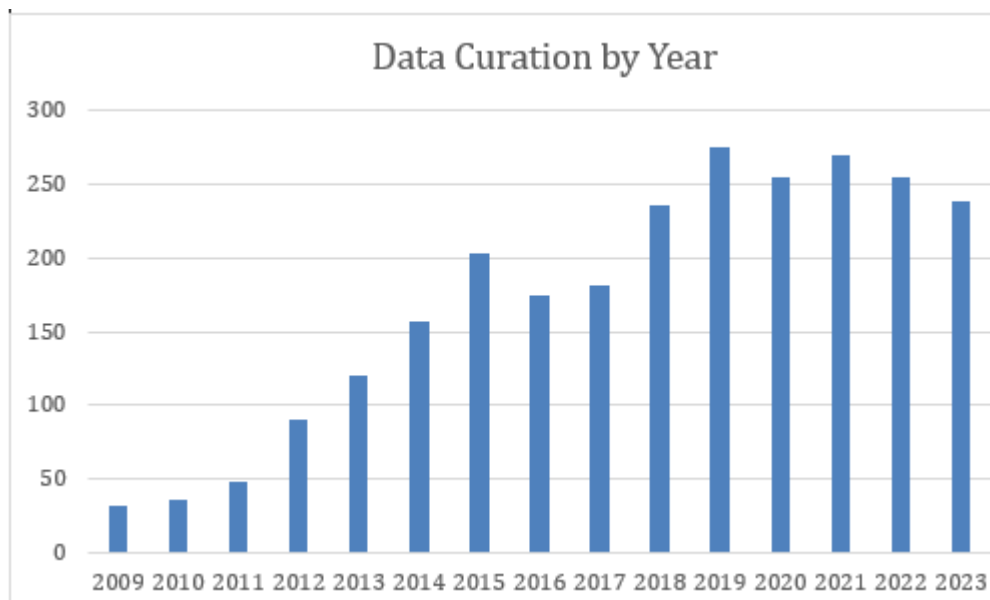
Scopus FAIR Principles Publication Trend by Year



Scopus was employed again to search the subject of “Data Curation,” using the search parameters “Data Curation” and “Digital Curation.” It produced a total of 2761 published articles. The years from 1995 to 2008 averaged seven published articles. From 2009 to 2023, that figure jumped to 168. Figure 5 shows how this topic also became increasingly relevant.

Figure 5

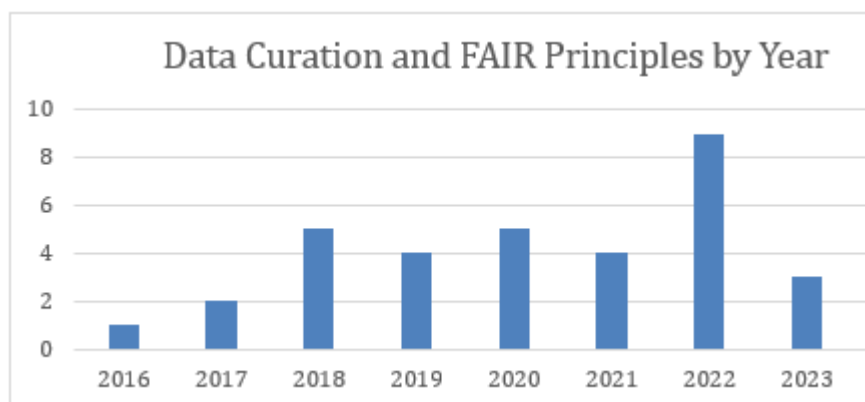
Scopus Data Curation Publication Trend by Year



Combining search parameters “Data Curation” and “FAIR Principles” led to a total of 36 articles from 2016 to 2023 an average of four articles per year, as shown in Figure 6.

Figure 6

Scopus Data Curation and FAIR Principles Publication Trend by Year



An analysis of the Scopus results indicates a clear and increasing trend of publication rates in RDM, FAIR Principles, and data curation, underscoring their growing importance in academia and research. The substantial rise in interest in RDM, beginning in 2012, likely

stemmed from heightened recognition of the need for effective research data management. It was likely bolstered by policies from organizations such as the National Institute of Health (NIH) and the NSF, which mandated open research and RDM Plans. Similarly, the increasing number of publications on FAIR Principles indicates a rising emphasis on making data Findable, Accessible, Interoperable, and Reusable, in response to the demands of open science and open data initiatives (Arpin & Kambesis, 2022; Jacobsen et al., 2020; Mons, 2018).

Despite the upward trends in these areas, a literature gap exists at the intersection of FAIR Principles and data curation. Scopus search results show that only four academic articles address the integration of these two topics. This gap highlights the need for further research to explore and clarify the synergy between data curation practices and adherence to FAIR Principles.

By investigating the combined application of FAIR Principles and data curation, this research will provide insights and contribute to the development of strategies for effective data management in the research community. This thesis aims to bridge the existing gap, offering an analysis and actionable recommendations to enhance the integration of these critical areas, advancing the field of RDM and supporting the goals of open science.

Scope and Organization

The literature review of this thesis will encompass several interconnected areas within the domain of RDM and open science. It will begin by exploring RDM, defining its core concepts and emphasizing its role in the modern research ecosystem. This foundation will lead to an examination of open science and open data, explaining their definitions and exploring various open data initiatives. The review will assess the benefits and challenges associated with open data in research, providing a perspective on its impact on scientific progress.

Building on these broader concepts, the literature review will investigate the nuanced definitions of data and research data, analyzing their similarities and differences. It will further refine this discussion by examining the characteristics of datasets, distinguishing between open and traditional datasets and their respective roles in scientific inquiry. A portion of the review will include exploring key frameworks for RDM, including the FAIR Principles, CARE Principles, and TRUST Principles. This section will describe these frameworks and evaluate their benefits and challenges in practical application.

The review will then transition to data curation, defining its processes and highlighting its significance in the context of research management. Economic considerations will also be addressed, exploring how they impact RDM practices and discussing the potential benefits economic policies could bring to open data and open science initiatives. The literature review will also examine the challenges inherent in implementing RDM best practices, emphasizing the importance of standardized methods and metadata. The review will conclude with an exploration of metadata's role in enhancing the FAIR compliance of datasets, comparing and contrasting its application in open datasets versus traditional ones. This comprehensive approach will provide an understanding of the current state and future directions of RDM and open science.

Introduction to Research Data Management

The concept of managing research data systematically can be traced back to the early 20th century, with roots in the scientific method and the need for accurate record-keeping. During this period, researchers began to recognize the importance of maintaining detailed and organized records of their experiments and observations. This recognition was driven by the growing complexity of scientific research and the need for reproducibility in scientific findings (Miyakawa, 2020; Resnik & Shamoo, 2017).

One of the earliest formal discussions of RDM can be found in the work of Karl Pearson, a prominent statistician and biometrician. In his 1892 book *The Grammar of Science*, Pearson emphasized the importance of careful data collection and analysis in scientific research (Pearson, 1957). Similarly, Ronald Fisher's work in the 1920s and 1930s on experimental design highlighted the need for systematic approaches to data collection and analysis in agricultural research (Fisher, 1936).

Despite these early recognitions, formalized RDM practices were not widespread during this period. Data management was often limited to individual research practices, which varied greatly in terms of methodology. The lack of standardized approaches to data management meant that valuable research data was often lost or became unusable over time. It wasn't until the latter half of the 20th century, with the advent of digital technologies and the exponential growth in data volume, that more structured approaches to RDM began to emerge (Borgman, 2012; Tolle et al., 2011).

Definition and Importance of RDM

RDM is the management of research data. Whyte and Tedds (2011) suggest that RDM concerns the organization of data, from its entry into the research cycle through to the dissemination and archiving of valuable results (Whyte & Tedds, 2011). Zhang says the goal of RDM is to ensure the availability, authenticity, and validity of scientific research (X.-F. Zhang, 2021). Andrikopoulou et al. maintain that RDM is comprised of many processes described as digital curation. It is the active management of research data that reduces the threats to long-term research value and reduces the risk of digital obsolescence (Andrikopoulou et al., 2021).

RDM has become increasingly important in the modern research landscape due to its role in enhancing the quality, efficiency, and impact of scientific research. The importance of RDM

can be understood through several key aspects. Firstly, RDM ensures integrity and reproducibility of research findings. As Wilkinson et al. argue, proper data management practices enable researchers to verify and build upon existing research, which is fundamental to the scientific method. By maintaining well-organized and properly documented data, researchers can more easily replicate and validate results, thereby increasing the reliability of scientific knowledge (Wilkinson et al., 2016). Azeroual et al., agree with Wilkinson et al.—that RDM is becoming more important to enable and ensure that research results can be verified and interpreted. Additionally, they agree that research results are made usable and actionable (Azeroual et al., 2022).

Secondly, RDM facilitates data sharing and collaboration among researchers. Borgman emphasizes that effective data management practices enable researchers to share their data more easily, fostering collaboration and accelerating scientific progress. This is particularly important in an era of increasing data-intensive and interdisciplinary research, where the ability to access and integrate diverse datasets can lead to new insights and discoveries (Borgman, 2012).

Azeroual et al. wrote in the article “Putting FAIR Principles in the Context of Research Information: FAIRness for CRIS and CRIS for FAIRness” that due to a growing volume of data and the cross-disciplinary challenges, there is a need for a data to be accessible and reusable (Azeroual et al., 2022).

Thirdly, RDM plays a role in maximizing the value and longevity of research data. Corti et al. emphasize that proper data management ensures that data remains accessible and usable over time, even as technologies and formats evolve. This long-term data preservation is important to build upon previous research, enable longitudinal studies, and facilitate reproducibility in science (Corti et al., 2019). Ailamaki et al., discuss the importance of data

preservation and the challenges associated with data versioning. They argue that data should be stored in repositories that support versioning capabilities, allowing for the archiving of original data while maintaining records of subsequent updates. This approach enables researchers to track changes over time and refer to specific versions of datasets. The authors propose that Relational Database Management Systems (RDBMS) are the optimal technological choice for managing scientific data, offering advantages such as structured storage of complex datasets, efficient querying and retrieval, built-in support for data integrity and consistency, and scalability to handle large volumes of research data (Ailamaki et al., 2010).

Furthermore, RDM is essential for compliance with funding agency requirements and ethical standards. Many funding bodies now mandate data management plans as part of grant applications, recognizing the importance of proper data stewardship (Fecher et al., 2015). Bloemers and Montesanti in their article “The FAIR Funding Model: Providing a Framework for Research Funders to Drive the Transition Toward FAIR Data Management and Stewardship Practices”, discuss how the research funding organizations (RFOs) are requiring grant holders to deliver reusable data as output from their research projects, and to share their data to contribute to future research. They propose a model that improves on the practices for RDM, including using data management plans (DPS) and the FAIR Principles framework (Bloemers & Montesanti, 2020). Additionally, RDM practices help ensure the ethical use of data, particularly when dealing with sensitive or personal information (Donaldson & Koepke, 2022).

Lastly, effective RDM can lead to increased visibility and impact of research. Piwowar and Vision found that studies that made their data openly available received more citations than similar ones that did not, suggesting that good data management practices can enhance the impact and recognition of research outputs (Piwowar & Vision, 2013). Tenopir et al. discuss the

impact proper RDM practices have on higher citation rates and broader impact across disciplines. They attribute this to having well-managed and accessible data that enables meta-analysis and systematic reviews (Tenopir et al., 2011).

Challenges in RDM

The phenomenon of having excessive data, commonly referred to as “data overload” or “information overload,” significantly impacts data manageability. According to Mahdi et al., information overload—considering its numerous dimensions, decentralized nature, and more dispersed details—makes useful information harder to find (Mahdi et al., 2020).

As the volume of available data increases, it becomes increasingly challenging to manage and organize it effectively. This can lead to data being stored in disparate locations such as data repositories, leaving datasets to be “siloeed.” A researcher searching for marine debris data, for instance, may have to visit numerous repositories to find a suitable dataset. Data formats can exist in many different variations, and some are not conducive to analysis. These include datasets saved as PDF files, or those within badly designed dataset containers. Many data repositories use different IT systems to store data. One database may contain an entire dataset, for example, while another—a data catalog—contains only a pointer to the dataset location. Varied location, format, and IT systems can encumber location of specific datasets. A study by Gantz and Reinsel on the digitalization of the global economy highlights the exponential growth of data and the challenges it poses for data management (Reinsel et al., 2018).

An overabundance of data can make search mechanisms less efficient. Searchers can yield too many results, many of which may be irrelevant or of low quality, making it difficult to find the most pertinent datasets. A lack of FAIR compliance can be another hindrance in a dataset. The phenomenon of “search cost” in information retrieval, as discussed by Hjørland and

Wilson, illustrates how increased data can complicate the process of finding the right information. Liu et al. define “search cost” as the behavioral efforts expended by users to carry out their search, and to understand results that may not return desired search goals (and, in fact, may diminish the rate of information gains) (Liu et al., 2016).

In a vast sea of data, individual datasets become less visible. This is particularly true for older or less frequently cited data, which can be buried under newer or more popular datasets. The concept of the “long tail” in data, as discussed by Anderson, explains how many less popular items can collectively have a significant presence but individually remain obscure (Anderson, 2006). The “long tail” in science often refers to the vast number of smaller, specialized, and often underrepresented scientific studies and datasets that do not receive as much attention as major projects. Collectively, however, they represent a significant portion of scientific data (B. Zhang et al., 2016).

Proper metadata and indexing are crucial for data findability, reuse, and interoperability. With data amounts often overwhelming, creating accurate and comprehensive metadata for each dataset becomes a critical, albeit formidable task. This was highlighted by Lynch in his discussion of the importance of metadata in digital curation (Lynch, 2008). Tsueng et al. maintain that findability requires useful and complete descriptions of data contents, which they define as metadata. This metadata allows researchers to discover and evaluate whether the dataset is suited for their purposes. Tsueng et al., also note that metadata standardization, such as schema.org., improves the management of data. However, a lack of metadata standardization thwarts efforts to combine separate data resources into a single, searchable index (Tsueng et al., 2023).

Data Management Lifecycle Models

As mentioned in previous sections, RDM is the management of research data. Whyte and Tedds describe RDM as encompassing the entire lifecycle of data within the research process. According to their definition, it involves organizing data from the moment it enters the research cycle, continuing through the various stages of analysis and use, and ultimately concluding with the dissemination and archiving of valuable results. This comprehensive view emphasizes that effective data management is not a single action but a continuous process that spans the duration of a research project and beyond (Whyte & Tedds, 2011). The RDM lifecycle model offers a systematic framework for understanding and planning the various operations required to manage data effectively throughout its entire lifespan (Ball, 2012; Qin et al., 2017).

As Ball and others point out, there are many different RDM lifecycle models, but all have common elements. Some RDM lifecycle models and their activities are shown in Table 1 below.

Table 1

RDM Lifecycle Models

Lifecycle Model	Lifecycle Elements
DCC Curation Lifecycle	Create or Receive, Appraise & Select, Ingest, Preservation Action, Store, Access, Use & Reuse, Transform
DDI Combined Lifecycle	Study Concept, Data Collection, Data Processing, Data Distribution, Data Discovery, Data Analysis, Repurposing, Data Archiving
ANDS Data Sharing Verbs	Create, Store, Describe, Identify, Register, Discover, Access, Exploit
DataONE Data Lifecycle	Plan, Collect, Assure, Describe, Preserve, Discover, Integrate, Analyze

Capability Maturity	Data acquisition, processing and quality assurance, Data description and representation, Data dissemination, Repository services/preservation
---------------------	---

Note: Adapted from "*Review of data management lifecycle models*" by A. Ball, 2012, University of Bath. (<https://purehost.bath.ac.uk/ws/portalfiles/portal/206543/redm1rep120110ab10.pdf>). In the public domain.

All models include stages for data creation/collection, description, storage, and discovery/access. Most also include some form of data analysis or use. The differences include that the DCC model emphasizes curation and preservation more than others. The DDI model includes a “Study Concept” stage, which is unique among these models. The ANDS model uses action verbs, focusing on the activities, rather than stages. The DataOne model includes a specific “Plan” stage, which is not explicit in the others (ARDC, 2024; Ball, 2012; Borghi et al., 2018; Qin et al., 2017).

The focus of each model is different, according to its purpose. For example, the DCC model focuses on curation and preservation. The DDI model is more focused on the research data lifecycle, while the ANDS model concentrates on data-sharing actions. The Capability Maturity Model (CMM) emphasizes organizational maturity when categorizing specific activities of the RDM lifecycle model. The DataONE model features the most comprehensive RDM processes.

These models represent different perspectives on managing research data throughout its lifecycle. While they share many elements, each has a unique focus and level of detail. Model selection would depend on the specific needs of the research project or organization, with some

more suitable for data curation (DCC), others for comprehensive research management (DataONE), and still others for emphasizing data sharing (ANDS) (Ball, 2012; Eaker, 2016).

Of the five RDM lifecycle models listed, only DataONE's holistic approach covers data sharing as well as critical aspects of planning, quality assurance, preservation, and reuse. These elements are crucial for effective open science practices (Michener & Jones, 2012).

Open Science and Open Data

The concept of open science is to make scientific research, data, and dissemination accessible to all levels of society. It encompasses various practices and principles designed to make the scientific process more transparent, collaborative, and accessible (Allen & Hartland, 2018). Vicente-Saez and Martinez-Fuentes believe that open science involves making scientific knowledge (publications, data, physical samples, and software) freely available under terms that enable reuse, redistribution, and reproduction of the research and its underlying data and methods (Vicente-Saez & Martinez-Fuentes, 2018).

Open science represents a transformative paradigm shift in the way scientific research is conducted, shared, and used across society. This movement advocates for transparency, collaboration, and accessibility at every stage of the scientific process, from initial inquiry to final dissemination. By breaking down traditional barriers to knowledge, open science aims to democratize access to scientific information, accelerate innovation, and enhance the reliability and reproducibility of research findings. This approach encompasses various practices, such as open-access publishing, open data sharing, open peer review, and citizen science initiatives (Spicer, 2018).

As Fecher and Friesike believe, open science is not just about making research outputs freely available, but about reimagining the entire scientific ecosystem to be more inclusive,

efficient, and impactful (Fecher & Friesike, 2014). The principles of open science align closely with the FAIR data principles (Findable, Accessible, Interoperable, and Reusable), as outlined by Wilkinson et al., which provide a framework for optimizing the reuse of scientific data (Wilkinson et al., 2016). As emphasized by Vicente-Saez and Martinez-Fuentes, open science has the potential to foster greater public trust in scientific institutions and findings by increasing transparency and public engagement in the scientific process (Vicente-Saez & Martinez-Fuentes, 2018). This paradigm shift challenges longstanding academic traditions and publishing models, pushing for a more collaborative and open approach to knowledge creation and dissemination that can benefit researchers, institutions, and society at large (Arzberger et al., 2004; Nosek et al., 2015).

Open Data Initiatives and Policies

Open data initiatives have become increasingly prominent as governments, organizations, and academic institutions recognize the value of making data accessible to the public. These initiatives aim to enhance transparency, foster innovation, and promote economic growth by enabling the free flow of information. One notable example is the Open Government Data (OGD) movement, which has been adopted by many countries worldwide. The OGD goal is to provide the public with access to government-held data to improve public services, increase government accountability, and spur innovation in various sectors (Ubaldi, 2013). Studies have shown that OGD initiatives can lead to significant societal benefits, including increased civic engagement, improved service delivery, and the creation of new business opportunities (Charalabidis et al., 2018; Janssen et al., 2012).

Academic institutions have also played a crucial role in promoting open data policies. The open access movement in academia advocates for free and unrestricted access to research

outputs, including datasets, to enhance the dissemination and impact of scientific knowledge. The FAIR Principles, developed by Wilkinson et.al, provide a framework for managing and sharing research data. The principles aim to ensure that data is well-documented, easily discoverable, and usable by both humans and machines, maximizing its potential for reuse and reproducibility (Wilkinson et al., 2016). The adoption of FAIR Principles has been supported by numerous funding agencies and research organizations, which mandate open data policies as a condition for grant funding (Mons et al., 2017).

In addition to FAIR, the CARE Principles (Collective Benefit, Authority to Control, Responsibility, and Ethics) have emerged to address the ethical and equitable use of data, particularly concerning Indigenous data sovereignty. Developed by the Global Indigenous Data Alliance (GIDA), the CARE Principles emphasize the need to respect Indigenous rights to control their data and ensure that data practices benefit Indigenous communities. This framework highlights the importance of ethical considerations in open data policies and aims to balance the benefits of data sharing with the need to protect the rights and interests of marginalized communities (Carroll et al., 2020).

Moreover, the TRUST (Transparency, Responsibility, User Focus, Sustainability, and Technology) principles provide guidelines for creating and maintaining trustworthy digital repositories. These principles focus on ensuring that digital repositories operate transparently, take responsibility for data stewardship, prioritize user needs, maintain sustainability, and employ robust technologies to safeguard data integrity. The implementation of TRUST principles is vital for building and maintaining public trust in open data initiatives, as they provide assurance that data will be managed responsibly and preserved for future use (Lin et al., 2020).

Overall, open data initiatives and policies are critical for promoting transparency, fostering innovation, and ensuring the ethical use of data. The adoption of frameworks such as FAIR, CARE, and TRUST principles underscores the importance of managing and sharing data in ways that maximize its value while protecting the rights and interests of all stakeholders (Dunning et al., 2017; Stall et al., 2019).

Benefits and Challenges of Open Data in Research

Open science promotes transparency, accessibility, and collaboration. Its benefits include improving the reproducibility of studies, fostering innovation through shared data and methodologies, and accelerating scientific discoveries by breaking down traditional barriers. However, challenges persist, such as ensuring data privacy, securing funding for open-access platforms, and addressing the lack of standardized protocols across disciplines (Fecher & Friesike, 2014; Vicente-Saez & Martinez-Fuentes, 2018).

Challenges of Open Data

Open data, while offering benefits, presents challenges that must be addressed to realize its full potential. Overcoming these challenges requires effort from governments, institutions, and the private sector to establish clear guidelines and promote a culture of data literacy and ethical usage. The challenges include data management and organization, search efficiency, information overload, lack of metadata standards, and resource constraints.

Difficulty in Data Management and Organization: As the volume of available data increases, it becomes increasingly challenging to manage and organize it effectively. This can lead to data being stored in disparate locations such as data repositories, leaving datasets to be “siloeed” (Mons et al., 2017). A researcher searching for marine debris data, for instance, may have to visit numerous repositories to find a suitable dataset. Data formats can exist in many

different variations, and some are not conducive to analysis. These include datasets saved as PDF files, or those within badly designed dataset containers. Many data repositories use different IT systems to store data. One database may contain an entire dataset, for example, while another—a data catalog—contains only a pointer to the dataset location. Varied location, format, and IT systems can encumber location of specific data sets (Labadie et al., 2020). A study by Gantz and Reinsel on the digitalization of the global economy highlights the exponential growth of data and the challenges it poses for data management (Reinsel et al., 2018).

Compromised Search Efficiency: An abundance of data can make search mechanisms less efficient. Searchers can yield too many results, many of which may be irrelevant or of low quality, making it difficult to find the most pertinent datasets. Another aspect can be the lack of FAIR compliance present in the dataset (Azeroual et al., 2022). The phenomenon of “search cost” in information retrieval, as discussed by Hjørland and Wilson (Hjørland & Wilson, 1997), illustrates how increased data can complicate the process of finding the right information. Liu et al. define “search cost” as the behavioral efforts expended by users to carry out their search, and to understand results that may not return desired search goals (and, in fact, may diminish the rate of information gains) (Liu et al., 2016).

Cognitive Overload: In a vast sea of data, individual datasets become less visible. This is particularly true for older or less frequently cited data, which can be buried under newer or more popular datasets. The concept of the “long tail” in data, as discussed by Anderson (Anderson, 2006), explains how many less popular items can collectively have a significant presence but individually remain obscure. The “long tail” in science often refers to the vast number of smaller, specialized, and often underrepresented scientific studies and datasets that do not receive as much attention as major projects. Collectively, however, they represent a

significant portion of scientific data (B. Zhang et al., 2016).

Challenges in Metadata and Indexing: Proper metadata and indexing are crucial for data findability, reuse, and interoperability. With data amounts often overwhelming, creating accurate and comprehensive metadata for each dataset becomes a critical, albeit formidable task. This was highlighted by Lynch in his discussion of the importance of metadata in digital curation (Lynch, 2008). Tsueng et al. maintain that findability requires useful and complete descriptions of data contents, which they define as metadata. This metadata allows researchers to discover and evaluate whether the dataset is suited for their purposes. Tsueng et al., also note that metadata standardization, such as schema.org, improves the FAIR compliance of datasets on the web. However, a lack of metadata standardization thwarts efforts to combine separate data resources into a single, searchable index (Azeroual et al., 2022; Tsueng et al., 2023).

Resource Constraints: Resources for data curation, including human expertise and technological tools, are often limited. With an ever-growing volume of data, these resources can become stretched thin, handicapping the quality of data curation and, consequently, discoverability, as noted by Palmer et al. In their article, “Scholarly Information Practices in the Online Environment Themes from the Literature and Implications for Library Service Development,” the authors discussed that resource constraints can include financial limitations due to limited budgets. In turn, these negatively impact data curation technology and infrastructure. Additional resource constraints include people; proper data management and curation requires skilled personnel, and these professionals are difficult to recruit and retain (Palmer, 2009; Tenopir et al., 2011).

Additionally, the rapid pace of change can constrain technology resources, which requires continuous investment: system updates, data curation tools, etc. Other resource

constraints may be the physical location of repositories and the ongoing need for infrastructure in digital data storage. Moreover, access to certain technologies may be challenging (Stodden et al., 2013; Tolle et al., 2011).

Benefits of Open Data

Open data offers a multitude of benefits that can drive innovation, transparency, and economic growth. By making data freely available to the public, it enables researchers, businesses, and policy makers to access valuable information that can inform decision-making. These benefits include improved reproducibility, increased collaboration, cost efficiencies, and greater accessibility.

Enhanced Reproducibility: Open data practices allow researchers to share their data, methods, and results openly, enabling other scientists to replicate studies and verify findings. This transparency helps to reduce errors and fraudulent practices, thereby strengthening the reliability of scientific research (Nosek et al., 2015). McKiernan et al., believe that open research practices benefit the researchers by gaining higher citation rates, attracting additional potential collaborators, and increasing funding opportunities due to gaining recognition and improving available resources (McKiernan et al., 2016). Wilkinson et al., support open data practices by ensuring that data are not only open but also usable and valuable for replication and verification by other researchers. Their goal in their FAIR Principles framework is to support open data by improving the overall efficiency and impact of scientific research through improved data management and sharing (Wilkinson et al., 2016).

Increased Collaboration: By making research outputs accessible to a broader audience, open data fosters collaboration across disciplines and geographical boundaries. This openness can lead to new insights and innovative approaches that might not emerge in a more closed

research environment (Vicente-Saez & Martinez-Fuentes, 2018). Molloy emphasizes that open data fosters collaboration by making research outputs accessible. This openness encourages cross-disciplinary research and the sharing of diverse perspectives, leading to innovative approaches and solutions that may not arise in more closed environments (Molloy, 2011). Fry et al., agrees with both Vicente-Saez & Martinez-Fuentes in that open data and shared research methods facilitate cross-disciplinary synergy. By making data and methodologies accessible across various fields, researchers from different disciplines can collaborate more easily. The ability to combine diverse perspectives enhances the depth of research outcomes (Fry et al., 2009).

Accelerated Discovery: Open access to research data and publications can speed up the pace of scientific discovery. Researchers can build on existing work without delays, leading to faster advancements and the development of new technologies and treatments (Piwowar et al., 2018). Molloy agrees with Piwowar et al., and points out that open data allows for the reuse of data in new and unforeseen ways, which can lead to novel insights and advancements and that by sharing data openly, researchers can build upon existing work more efficiently and accelerate the pace of discovery (Molloy, 2011).

Greater Accessibility: Open data provides greater access to scientific knowledge, allowing not only scientists but also policymakers, educators, and the public to benefit from research findings. This widespread accessibility can inform decision-making (Tennant et al., 2016). Molloy argues that open data practices allow for a wider audience, including the general public, to access and benefit from scientific research, promoting greater transparency (Molloy, 2011). McKiernan agrees with both Tennant et al., and Molloy that open data facilitates broader access to scientific information, benefiting various stakeholders (McKiernan et al., 2016).

Cost Efficiency: Sharing resources and data openly can reduce duplication of effort and save costs associated with data collection and research. Open data practices promote the efficient use of funding and resources, ensuring that research investments yield maximum benefits (Molloy, 2011). Fry et al., promote the reuse and sharing of data, and that when data is readily available, researchers can avoid unnecessary duplication of data collection efforts, saving time and resources. This efficiency allows for more resources to be allocated to analysis, interpretation, and new data collection in under-researched areas, thereby broadening the scope of scientific inquiry (Fry et al., 2009).

Perspectives on What Constitutes Research Data

The importance of the data itself should not be understated. It is a foundational component of research and decision-making processes, yet it is a multifaceted concept with diverse interpretations across disciplines. According to Lawal, data collection and the subsequent data allows scientists to test hypotheses through controlled methods and statistical analysis. They are important for confirming or refuting scientific theories. Data serves as the cornerstone of scientific research, providing a foundation upon which new discoveries and advancements are constructed (Lawal, 2010). Data supports the establishment of facts and findings in science, ensuring that conclusions are based on observable and verifiable results. The collection, analysis, and interpretation of data enable scientists to test hypotheses, validate existing theories, and forge new knowledge across a range of disciplines. Verbaan and Cox write that by using data analysis, researchers can explore new areas of inquiry, solve complex problems, and contribute to technological and medical advancements (Andrikopoulou et al., 2021; Verbaan & Cox, 2014).

Tim Berners-Lee is quoted as saying “Data is a precious thing and will last longer than the systems themselves” (Chancellor, 2020). Stephen Humphreys considers data to be the

material nature of the binary digit as the basic unit of contemporary data. He further believes data is reconstituted information (Humphreys, 2018). Buckland states that an artifact or observation may be at best “alleged evidence” (Borgman, 2012; Buckland, 1991). Authors Gerald van Belle and Leslie Ruiter consider data in any form as bare, naked facts, and that data, information, and knowledge are frequently used for concepts that, in fact, overlap. They say that data, on its own, carries no meaning. For data to become “information,” it must be interpreted and take on a meaning (van Belle & Ruiter, 2014). Data is an important asset for any organization and should be recognized and managed as such (Arpin & Kambesis, 2022).

In the paper “Data as Assemblage”, Ceilyn Boyd defines data by using the concept of assemblage. Data consists of forms and occurrences that are changeable and portable; they are sociotechnical arrangements of material (Boyd, 2022). Data itself can come in many forms: biospecimen, video recordings, images, software programs, algorithms, paper lab notebooks, etc. It is useful to think of data as everything that would be needed to reproduce a given scientific output, and that every data element has a story (Surkis & Read, 2015).

In their article “Making Research Data Accessible”, Kapiszewski and Karcher note that definitions of “data” or “research data” vary widely, with some emphasizing content, such as the National Research Council’s definition: “Data are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors.” Others, such as the National Academy of Science, define data by its use. Consider: “information used in scientific, engineering, and medical research as inputs to generate research conclusions.” Data serves as the empirical building blocks of knowledge, differing across disciplines based on research goals and methods. What sets data apart from general information is its transformation into forms suitable

for measurement and analysis, making it essential for knowledge generation (Kapiszewski & Karcher, 2020; Lawal, 2010).

Building on the foundational understanding of data, it is important to explore the specifics of research data in greater detail (Borgman, 2012). Research data can be defined as information in documents or digital form collected, observed, created, or produced during scientific research. It is also evidence derived during the research process, or information commonly accepted in the research community required to validate research findings and results (Elsevier, 2022; Imperial College of London, 2022).

Research data (also known as scientific data or scholarly data) is an essential artifact of scientific work: it leads to insights, makes research reproducible, and validates findings. OpenAIREplus, a European Open Access Infrastructure for research, defines research data as any kind of data produced during scientific research. This includes databases of raw data, tables, and pictures (Guibault & Wiebe, 2013). Research data can come in various forms and from multiple sources, each suited to different types of inquiries and disciplines. The most common forms of research data are quantitative, qualitative, and mixed methods (Creswell & Poth, 2024). Hanson et al. highlight the core responsibilities of the scientific community as promoting transparency, enforcing standardization, and ensuring the proper archiving of research data (Hanson et al., 2011). Research data offers one thing above all: the opportunity to generate valuable knowledge (Azeroual, 2020).

RDM is an important aspect of the scientific process that involves the organization, storage, preservation, and sharing of data generated during research activities. It is the management of research data that presents challenges for maintaining quality, integrity, and utility of research data. Proper RDM practices are crucial for ensuring the reliability,

accessibility, and reproducibility of research findings, which are essential components of the scientific method. As the volume and complexity of data continue to grow, implementation of robust data practices becomes increasingly important (Mons, 2018; Wilkinson et al., 2016).

Specific Characteristics of Research Data

Research data can be characterized by several key attributes that define its quality, usability, and value in scientific research. Understanding these characteristics is important for effective data management and utilization. The attributes include accuracy, completeness, consistency, timeliness, validity, accessibility, and reproducibility (Borgman, 2012; Carlson et al., 2011; Tenopir et al., 2011).

Accuracy: Accuracy refers to the correctness and precision of the data. It is crucial that research data accurately reflects the phenomena being studied to ensure valid conclusions. Inaccurate data can lead to erroneous results and interpretations (Pipino et al., 2002). Van den Broeck et al. emphasize that data accuracy is fundamental to the validity of research findings and that the cleaning of data is essential to accuracy and data integrity (Broeck et al., 2005).

Completeness: Completeness pertains to the extent to which all necessary data is present. Incomplete data can compromise the integrity of research findings by omitting critical information. Ensuring that datasets are comprehensive is vital for robust analysis and reliable outcomes (Kahn et al., 2015). Wang and Strong discuss the concept of data completeness as a dimension of data quality. They also define it as the extent to which all data is complete and the importance of data being contextually appropriate and accessible for the tasks at hand (Wang & Strong, 1996).

Consistency: Consistency involves maintaining uniformity and coherence across datasets. This characteristic ensures that data is presented in a standardized format, which

facilitates comparison and integration with other data sources. Inconsistencies can lead to confusion and misinterpretation (Batini et al., 2009). Pipino et al. consider data consistency as a dimension of data quality. They think consistent data is needed for accurate data analysis and decision making, as it helps to avoid misunderstandings and misinterpretations that may result from inconsistent data representations. Their solution to ensuring consistency involves applying standardized data formats, validation rules, and systematic data cleaning processes across the datasets (Pipino et al., 2002).

Timeliness: Timeliness refers to the currency and relevance of data. Research data should be up to date to provide accurate insights into current conditions or phenomena. Outdated data may not accurately represent the present state, leading to obsolete or irrelevant conclusions (Batini et al., 2009). Wang and Strong agree with Batini et al., in that data that is not timely can lead to incorrect conclusions (Wang & Strong, 1996).

Validity: Validity indicates that the data measures what it is intended to measure. This characteristic ensures that the data is appropriate for the research questions being addressed and that the methods of data collection and analysis are sound (Mellinger & Hanson, 2020). Fink emphasizes the importance of validity in her book *Conducting Research Literature Reviews: From the Internet to the Paper*. She highlights that without validity, the research findings could be misleading or incorrect (Fink, 2019).

Accessibility: Accessibility is the ease with which data can be obtained and used by authorized users. Data should be stored in a manner that allows for easy retrieval and use, which is essential for enabling verification, replication, and further research (Stevens, 2016). According to Borgman, accessible data must exist within a robust knowledge infrastructure – an ecosystem

of people, technologies, institutions, and practices that work together to manage and exploit data over the long term (Borgman, 2013).

Reproducibility: Reproducibility means that research data should be capable of being duplicated by others using the same methods. This characteristic is fundamental to the scientific method, as it allows for the verification of results and builds trust in the research findings (R. Peng, 2015). Stodden et al., agrees with Peng in that reproducibility is fundamental to scientific research and that to ensure reproducibility, the code and data needs to be accessible and in readable formats (Stodden et al., 2013).

Open Datasets

Open datasets have emerged as an important component of the open science movement, revolutionizing the way research is conducted and shared across various disciplines (Spicer, 2018). Wilkinson et al. introduced the FAIR Principles, which have become a cornerstone in the development and management of open datasets. These principles provide a framework for optimizing data reuse and have been widely adopted by data repositories and research institutions (Jacobsen et al., 2020; Mons, 2018; Wilkinson et al., 2016).

The importance of open datasets in advancing scientific research is well-documented in the literature. Pasquetto et al. highlight how open datasets serve two functions in the scientific process. Firstly, they facilitate reproducibility in science, allowing researchers to verify and validate their findings of their peers. Secondly, they enable meta-analysis, which can lead to new insights across various disciplines. By providing researchers with access to large, diverse datasets, open data initiatives foster innovative approaches to existing scientific questions and enable the exploration of new research avenues (Borgman et al., 2019; Pasquetto et al., 2017; Tenopir et al., 2011).

Supporting this perspective, Pampel and Dallmeier-Tiessen, in their chapter “The Vision of Open Research Data” from the book *Opening Science* (Fecher et al., 2014), further elaborate on the benefits of open datasets. They argue that open research data creates new possibilities for scientists by enabling them to re-use existing datasets in new ways. This not only promotes efficiency in research by reducing duplication of effort but also encourages creative applications of data that may not have been initially envisioned by the original researchers. Additionally, Pampel and Dallmeier-Tiessen emphasize that open datasets improve the verification process for scientific data. This enhancement in data verification ensures adherence to good scientific practices, promoting transparency and trust in the research process (Bartling & Friesike, 2014; R. Peng, 2015). This aligns with the broader goals of open science, as discussed by Fecher and Friesike, who emphasize the transformative potential of open data practices in scientific research (Fecher & Friesike, 2014).

In the realm of innovation and economic impact, Janssen et al. explore the benefits of open datasets, noting their potential to spur innovation in both academic and commercial sectors. Their work underscores how open data can lower barriers to entry for researchers and entrepreneurs, particularly benefitting those with limited resources (Janssen et al., 2012). The economic potential is further quantified by the European Data Portal, which estimated the market size of open data to be 184.45 billion EUR for the EU28+ in 2019 (Publications Office of the European Union., 2020).

Despite these benefits, several challenges persist in the open dataset landscape. Borgman reviews the complexities of data sharing, highlighting issues such as quality control, standardization, and the need for proper attribution (Borgman, 2012). Other challenges include data quality, privacy, and effective use of open data resources. Janssen et al., discuss the various

barriers to the adoption of open data practices including information quality and technical barriers (Janssen et al., 2012). Borgesius et al. in the article “Open Data, Privacy, and FAIR Information Principles: Towards a Balancing Framework” discuss the issues between open data and privacy. These challenges are pertinent as the volume and variety of open datasets continue to grow (Borgesius et al., 2015).

The role of open datasets in fostering citizen science has been explored by Bonney et al., who discuss how open data enables non-professionals to contribute and engage with scientific research (Bonney et al., 2014). This democratization of science aligns with the broader goals of open science, as conveyed by Vicente-Saez and Martinez-Fuentes, who emphasize the potential of open practices to increase public trust in scientific institutions (Vicente-Saez & Martinez-Fuentes, 2018).

The technological infrastructure supporting open datasets has also been a subject of study. Assante et al. examine how advancements in data storage, processing, and sharing technologies have facilitated the growth of open datasets. Their work highlights the role of specialized data repositories and cloud computing in making large-scale data sharing feasible (Assante et al., 2016). Amorim et al., also conclude that data repositories are best suited to provide the technological infrastructure needed to support open datasets. This includes support for metadata, search mechanisms and the large volumes of data and the ability to scale accordingly (Amorim et al., 2015).

As open datasets become more prevalent, the need for data literacy has gained attention. Koltay discusses the implications of open datasets for education, emphasizing the need for new skills in data management, analysis, and interpretation among researchers and information professionals (Koltay, 2017).

Government and institutional initiatives have played a significant role in promoting open datasets. The European Commission's Open Data Directive exemplifies how policy measures can increase the availability and use of public sector data, further driving the open science movement (European Union, 2019).

Traditional Datasets

Traditional research data and datasets are indispensable in the realms of science and academia. They serve as fundamental instruments for research, analysis, and the progression of knowledge. These datasets are distinguished by their systematic and disciplined approach to data collection, organization, and curation, which are essential to ensure data integrity and usability (Borgman, 2013; Palmer, 2009; Tenopir et al., 2011).

Traditional research datasets are typically characterized by a high degree of structure and organization. This involves arranging data in a format that is logical, consistent, and conducive to analysis, such as tables, databases, or structured files. The structured nature of these datasets facilitates efficient data processing and analysis, making it easier for researchers to extract meaningful insights (Borgman et al., 2019).

The data in traditional datasets is collected through methodical and often standardized techniques, ensuring the reliability and validity of the information. These methods vary according to the discipline. However, they often include controlled experiments, surveys, field observations, or systematic literature reviews. The method selected is crucial for the dataset's relevance and applicability to the intended research question (Tenopir et al., 2011).

While traditional research datasets are fundamental for scientific advancement, their accessibility is often restricted. Access may be limited due to privacy concerns, proprietary rights, or specific research agreements. However, there is a growing movement toward open

access, where such datasets are made more widely available under specific conditions, promoting greater transparency and collaboration in research (Tenopir et al., 2011).

Traditional datasets are vital for advancing scientific knowledge. They provide a solid empirical foundation upon which research hypotheses are assessed and new theories are developed. Their structured and reliable nature makes them particularly valuable for longitudinal studies, comparative analysis, and validation of previous research findings (Borgman et al., 2019).

The management and application of traditional research datasets are not without challenges. These include dealing with the increasing volume and complexity of data, ensuring interoperability between different data systems, and adapting to new technological advancements in data analysis. Furthermore, the evolving landscape of data sharing and open access presents both opportunities and challenges for the curation and utilization of traditional research datasets (Palmer, 2009).

Traditional research datasets offer numerous benefits to the scientific and academic communities and are integral to the advancement of knowledge and research. These research datasets are often collected and curated with a focus on reliability and validity. Their inherent methodical approach to data collection ensures that the data is accurate, consistent, and trustworthy, making it a reliable foundation for research and analysis (Borgman, 2012).

Traditional research datasets typically provide a depth and detail of information that is invaluable for in-depth research. The comprehensive nature of these datasets allows researchers to conduct thorough analysis and derive nuanced insights (Tenopir et al., 2011). These datasets are particularly useful for longitudinal studies, where data is collected over extended periods.

They provide a historical record that can be critical for understanding trends, changes, and long-term patterns (Renear & Palmer, 2009).

Traditional research datasets often adhere to standardization formats and methodologies, making it easier to compare data across different studies and disciplines. This standardization facilitates meta-analyses and systematic reviews (Borgman et al., 2019). The structured and detailed nature of traditional datasets supports the reproducibility of research. Additionally, access to original data allows other researchers to validate findings and replicate studies, which is fundamental to the scientific process (Molloy, 2011). These datasets can be tailored to the specific needs and methodologies of different fields of study, making them particularly valuable resources within those domains. They help build a body of knowledge that is specific and relevant to each discipline (Tenopir et al., 2011).

Traditional datasets are often curated with attention to ethical and legal considerations, especially when dealing with sensitive or confidential information. This ensures that the data is handled responsibly and in compliance with relevant laws and guidelines (Palmer, 2009). Molloy believes that traditional datasets provide a comprehensive and reliable data source, and reduce the need for redundant data collection, thereby saving time and resources. They enable researchers to build upon existing knowledge rather than starting from scratch (Molloy, 2011).

In conclusion, traditional research datasets are one of the cornerstones of scientific inquiry. They offer structured, dependable, and methodically collected data crucial for research across various disciplines. Despite facing challenges in the ever-evolving landscape of data management and technology, these datasets continue to be a fundamental resource for knowledge discovery and advancement. The benefits of traditional research datasets discussed in

this section included reliability, standardization, comparability, and reproducibility (Borgman, 2012; Douglass et al., 2014; Tenopir et al., 2011).

Frameworks for Research Data Management

RDM is an important component of modern scientific practice, playing a role in ensuring the value and usability of data throughout its lifecycle (Borghi et al., 2018). As the volume and complexity of research data continue to grow exponentially, effective RDM frameworks provide structured approaches to handle data from creation and collection to preservation and dissemination (Tenopir et al., 2011). These frameworks aim to address the challenges of data accessibility, reliability, and ethical management in an increasingly data-driven research landscape (Wilkinson et al., 2016).

In recent years, several frameworks have been developed to guide best practices in research data management. Among these are the FAIR, CARE, and TRUST principles. Each offers a unique perspective on how to maximize the utility and integrity of research data while respecting the rights of all stakeholders (Carroll et al., 2020; Lin et al., 2020; Wilkinson et al., 2016). These frameworks address various aspects of data management, including technical standards, ethical considerations, and organizational practices.

The development and adoption of these principles reflect the growing recognition of the importance of proper data management in advancing scientific knowledge and ensuring research reproducibility (G. Peng et al., 2021). It also highlights the evolving nature of data stewardship, which must adapt to new technological capabilities, ethical considerations, and the changing needs of the global research community (Borgman, 2012).

By providing comprehensive guidelines for data handling, these frameworks contribute to the creation of a more robust, transparent, and collaborative research ecosystem. They encourage

practices that enhance the immediate value of research data and ensure its long-term preservation and potential for reuse in future studies (Mayernik, 2016).

As the field of RDM continues to evolve, these principles serve as foundational elements in shaping policies, informing best practices, and guiding the development of data structures across various global scientific disciplines and institutions (European Commission., 2024).

FAIR Principles

Hanson et al. underscore the importance of widely disseminating data within the scientific community, noting the growing challenge to ensure that research data are adequately described, standardized, archived, and universally accessible (Hanson et al., 2011). Several frameworks have been created to address these challenges and enhance the usability of research data. For instance, M. Wilkinson and a diverse team from industry, academia, funding bodies, and scholarly publishing introduced the FAIR Principles in 2016. The principles provide a framework to guide best practices in RDM (Arpin & Kambesis, 2022; Bonino da Silva Santos et al., 2016; Wilkinson et al., 2016). These principles also provide a set of guidelines for managing research data in a way that maximizes its value and facilitates its use by the scientific community.

1. **Findable:** Data should be easily discoverable by both humans and machines. This can be achieved by using unique and persistent identifiers, such as Digital Object Identifiers (DOIs), and by providing rich metadata that describes the data.
2. **Accessible:** Data should be readily accessible to authorized users. This requires the implementation of clear access protocols and the use of open, standardized communication protocols.

3. **Interoperable:** Data should be able to be integrated with other data and be compatible with various applications and workflows. This can be achieved by using standardized data formats, vocabularies, and ontologies.
4. **Reusable:** Data should be well-documented and have clear usage licenses to enable its reuse by others. This includes providing detailed provenance information and using appropriate data licenses.

The FAIR Principles are an important key in the usability of research data. By making data Findable, Accessible, Interoperable, and Reusable, researchers can more easily build upon existing knowledge, collaborate with others, and accelerate scientific discovery. The principles also promote transparency and reproducibility in research, as they enable others to validate and verify research findings (Mons, 2018; Stall et al., 2019; Wilkinson et al., 2016).

CARE Principles for Indigenous Data Governance

Another framework is the CARE Principles for Indigenous Data Governance. The CARE Principles were developed by the Global Indigenous Data Alliance to complement the FAIR Principles and address the specific concerns and rights of Indigenous peoples regarding data governance. The CARE Principles stand for Collective Benefit, Authority to Control, Responsibility, and Ethics. These principles emphasize the importance of data governance, collective ownership, and the rights of Indigenous communities to control and derive benefit from data about their peoples, territories, and ways of life (Barsness et al., 2023; Carroll et al., 2020).

The CARE Principles describe high-level actions applicable within research, government, and institutional data settings. The goal is for data stewards and other users of Indigenous data to implement CARE and FAIR Principles in tandem. The CARE Principles are defined below and

have been adapted from Carroll et al. articles, “The CARE Principles for Indigenous Data Governance” and “Operationalizing the CARE and FAIR Principles for Indigenous Data Futures.”

1. Collective Benefits:

C1. For inclusive development and innovation: the benefits derived from the use of such data should contribute to the holistic well-being and self-determination of Indigenous communities and should not perpetuate existing inequalities or marginalization.

C2. For improved governance and citizen engagement: the collective benefits derived from the responsible use of Indigenous data should also contribute to improved governance and citizen engagement. This can include empowering Indigenous peoples to participate in decision-making processes that affect their lives, as well as enhancing transparency and accountability in governance structures.

C3. For equitable outcomes: the collective benefits should lead to more equitable outcomes for Indigenous peoples, addressing historical injustices and promoting social, economic, and environmental justice. This can involve using data to inform policies and programs that address disparities and inequalities faced by Indigenous communities.

2. Authority to Control:

A1. Recognizing rights and interests: the CARE Principles acknowledge the rights and interests of Indigenous peoples in relation to data about their communities, territories, and knowledge systems. This includes the right to self-determination and the right to control and govern data that is generated from or about their communities.

A2. Data for governance: Indigenous peoples have the authority to control and govern data about their communities, including how it is collected, used, and shared. This authority is rooted in their inherent rights to self-governance and self-determination.

A3. Governance of data: the CARE Principles recognize that Indigenous peoples have the right to establish their own data governance protocols and mechanisms, which should be respected and adhered to by external parties seeking to access or use data about their communities.

3. Responsibility:

R1. For positive relationships: the CARE Principles emphasize the importance of building and maintaining positive relationships between Indigenous peoples and those seeking to work with data about their communities. This involves fostering trust, respect, and reciprocity, as well as acknowledging and addressing historical injustices and power imbalances.

R2. For expanding capability and capacity: those working with Indigenous data have a responsibility to contribute to expanding the capability and capacity of Indigenous peoples to govern and manage their data. This can include providing training, resources, and support for developing data governance frameworks, data management systems, and data literacy.

R3. For Indigenous languages and worldviews: The care principles recognize the importance of preserving and promoting indigenous languages and worldviews in their context of data management in governance. This includes ensuring that data about Indigenous peoples is collected, stored, and presented in a way that reflects and respects their languages, cultural protocols, and knowledge systems.

4. Ethics:

E1. For minimizing harm and maximizing benefit: the ethical principle of the CARE framework emphasizes the importance of minimizing harm and maximizing benefits for Indigenous peoples when working with data about their communities. This involves conducting thorough risk assessments, implementing safeguards to protect privacy and confidentiality, and ensuring that the use of data aligns with the collective interest and well-being of Indigenous peoples.

E2. For Justice: The ethical principle also highlights the need for justice in the context of data management and governance. This includes addressing historical injustices, promoting equity and inclusion, and ensuring that the benefits derived from the use of Indigenous data contribute to the realization of social, economic, and environmental justice for Indigenous peoples.

(Barsness et al., 2023; Carroll et al., 2020, 2021).

TRUST Principles for Digital Repositories

A third framework for managing and sharing research data in an ethical, sustainable, and responsible manner is the TRUST Principles for Digital Repositories. The TRUST Principles were developed by a consortium of organizations to provide a set of guiding principles for digital repositories to ensure the long-term preservation and stewardship of digital data. The TRUST Principles stand for Transparency, Responsibility, User Focus, Sustainability, and Technology. These principles emphasize the importance of transparency in operations, responsible management of data, meeting user needs, sustainable funding, and governance models, and the user of appropriate technologies (Lin et al., 2020; G. Peng et al., 2021). The table below describes the TRUST Principles and their guidance for repositories.

Table 2*TRUST Principles for Digital Repositories*

Principle	Guidance for Repositories
Transparency	<p>To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.</p> <ul style="list-style-type: none">• Terms of use• Minimum digital preservation timeframe• Provide security for sensitive data
Responsibility	<p>To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.</p> <ul style="list-style-type: none">• Adhere to the designated community's metadata and curation standards• Stewardship of the data holdings• Provide data services such as data downloads or machine interfaces• Manage the intellectual property rights of data producers
Use Focus	<p>To ensure that the data management norms and expectations of target user communities are met.</p> <ul style="list-style-type: none">• Implement relevant data metrics and making them available to users• Providing community catalogs to facilitate data discovery• Respond to evolving community expectations
Sustainability	<p>To sustain services and preserve data holdings for the long-term.</p> <ul style="list-style-type: none">• Plan sufficiently for risk mitigation, business continuity, disaster recovery• Securing funding to enable ongoing usage and maintenance• Providing governance for long-term preservation of data so that data resources remain discoverable, accessible, and usable in the future
Technology	<p>To provide infrastructure and capabilities to support secure, persistent, and reliable services.</p> <ul style="list-style-type: none">• Implement relevant and appropriate standards, tools, and technologies for data management and curation• Have plans and mechanisms in place to prevent, detect, and respond to cyber or physical security threats

Note: Adapted from "*The TRUST principles for digital repositories*" by Lin et al., 2020 (<https://doi.org/10.1038/s41597-020-0486-7>). In the public domain.

Data Curation Definition and Importance

Data curation encompasses the management of data from its inception to ensure it remains reliable, accessible, and suitable for its designated uses. This involves not just the preservation of data, but also its active maintenance and enhancement, which may include continuous updates to keep dynamic data fit for purpose. The significance of data curation cannot be overstated, as it directly influences the availability and utility of data for the research community. Without proper curation, data may become obsolete, unreliable, or inaccessible, thereby affecting the research process and undermining the principles of open science. Through data curation, we ensure that research data adheres to high standards such as those established by the FAIR Principles, and remains a robust and valuable resource for scientific inquiry (Gonzalez & Peres-Neto, 2015; Lord et al., 2004; Sheridan et al., 2021).

Key Aspects of Data Curation

Data curation includes several multifaceted processes that encompass various activities beyond simply managing data for reliability. These include Data Organization, Data Quality, Metadata Creation, Archiving, Preservation, and Accessibility.

Data Organization: Data organization is a key component of data curation. It involves structuring and categorizing data in a logical, coherent manner, which facilitates navigation and retrieval. Data organization often includes classifying data into hierarchies, tagging, and using standard taxonomies, which are crucial for data management (Broman & Woo, 2018; Gilliland, 2008). Borgman argues that effective data organization is crucial for maintaining data integrity, enhancing accessibility, and maximizing the value of research data throughout its lifecycle

(Borgman, 2012). Gilliland also believes that well-organized data is essential for data integrity, improving discoverability, and supporting interoperability with other datasets and systems (Gilliland, 2008).

Data Quality: Data quality is another key aspect of curation. It includes processes to verify data accuracy, consistency, and completeness. High-quality data is essential for reliable research and decision-making. Sound data curation practices identify and rectify errors, inconsistencies, or gaps in datasets (Batini et al., 2009). Wang and Strong believe that poor quality data can have social and economic impacts. They emphasize their understanding that data should be high-quality, intrinsically good, clearly represented, and accessible to the data consumer (Wang & Strong, 1996).

Metadata Creation: Data curation also features detailed and accurate metadata. Metadata describes the data's origin, purpose, time of creation, creator, location, and format, among other details. Effective metadata enhances data discoverability while facilitating the understanding of data context and constraints. This is essential for its proper use (Duval et al., 2002; Palmer, 2009). The more highly structured an information object is, specifically datasets, the more that structure can be exploited for searching, manipulation, and interrelating with other datasets (Gilliland, 2008).

Archiving: Archiving involves storing data securely, and in ways that enable long-term retention. It ensures that data remains available for future research and reference, and safeguards against data loss due to technological obsolescence or other risks (Beagrie, 2006). In the article "Preserving Digital Materials: Confronting Tomorrow's Problems Today", Keene outlines several key principles and practices for effective digital archiving and underscores that digital

archiving is not a one-time process but an ongoing effort that requires continuous attention and resources to ensure the preservation and accessibility of digital information (Keene, 2002).

Preservation: Preservation is closely related to archiving. However, it focuses on maintaining data over time, ensuring it remains usable and accessible despite changes in technology and formats. This includes converting data into formats less likely to become obsolete and ensuring its physical and digital security (Keene, 2002). Conway highlights the unique challenges associated with digital preservation, including the rapid obsolescence of digital formats and technologies, the fragility of digital media, and the complexity of maintaining the integrity and authenticity of digital records over time (Conway, 2010). Rosenthal et al., agree with Conway in that the rapid pace of technology leads to digital formats and technologies becoming obsolete. The authors suggest that regular migration to current formats and technologies ensures ongoing accessibility (Rosenthal et al., 2005).

Accessibility: Accessibility ensures that data is available to its intended users. This involves creating and maintaining instinctive, easy-to-navigate user interfaces and access systems. It also includes efforts to make sure data is available to those who need it, while respecting privacy and ethical considerations (Jati et al., 2022). Data curation also involves ethical considerations, particularly in handling sensitive or confidential information. Compliance with legal and ethical standards is a must (Palmer, 2009). Borgman expands the legal and ethical standards by discussing the following key points: privacy and confidentiality, informed consent, intellectual property rights, and equitable access. Borgman emphasizes the importance of protecting the privacy and confidentiality of individuals whose data is included in research datasets. She points out that researchers and institutions must comply with legal requirements

such as data protection laws and ethical guidelines that safeguard personal information (Borgman, 2012).

The Role of Metadata

Metadata can have a significant impact on the compliance of FAIR Principles and in data curation processes on data. By providing information about the content, structure, and provenance of data, metadata can make it easier for users to find, understand, and use data (Arpin & Kambesis, 2022; Riley, 2017). Hauschke et al., state that metadata is a basic element of research information and should be FAIR-compliant. Metadata renders data FAIR-compliant, and is crucial in enabling findability, reusability, interpretation, and assessment of resources (Hauschke et al., 2021). Bloemers and Montesanti state that metadata management processes need key components such as RDM Planning, which involves creating Data Management Plans (DPMs) that comply with the FAIR Principles. The goal is for data to adhere to the FAIR Principles as much as possible, to ensure reliability. Some research-funding organizations (RFO) are also considering additional requirements to enhance FAIR compliance. These include promoting development and application of machine-actionable metadata or community-specific standards to improve research data interoperability (Bloemers & Montesanti, 2020).

Importance of Metadata for FAIR Compliance

In their article “Roadmap to FAIR Research Information in Open Infrastructures”, Hauschke et al., state that to find a data point, one must be able to index and discover it. Moreover, they declare that “metadata is the key to findability.” (Hauschke et al., 2021). Juty et al., emphasize the importance of persistent and resolvable identifiers in ensuring the reliability of large-scale data management across various domains and infrastructures. Well documented APIs for human- and machine-readable metadata, which include provenance

information as well as descriptive “guide” metadata, are crucial for FAIR compliance in a FAIR data ecosystem. Juty et al. emphasize the importance of robust links to provenance documentation for reliable data reuse. They also advocate sound data sharing, archiving, and citation practices, along with metadata procedures, to promote creation of FAIR data (Juty et al., 2020).

In the article “FAIR Principles: Interpretations and Implementation Considerations”, Jacobsen et al. provide an in-depth analysis of metadata within the context of the FAIR Principles. According to them, metadata encompasses any resource description that aids in FAIR compliance. They clarify the often-confused distinction between data and metadata, stating that within the FAIR framework, each data/metadata pair is treated distinctly. In this framework, metadata acts as a descriptor, while data is the entity being described, thereby ensuring clarity and avoiding ambiguity in their relations (Jacobsen et al., 2020).

Importance of Metadata in Data Curation

Metadata plays an important role in data curation processes by providing essential context and information about data, which enhances its usability and interoperability. Metadata describes various attributes of the data, such as its source, structure, and meaning, facilitating more effective data discovery and retrieval. This descriptive information is crucial for ensuring that data can be accurately interpreted and integrated with other datasets, which is necessary for research and analysis across different domains (Wilkinson et al., 2016). According to Ruggles, metadata in data curation is essential for efficient data integration and dissemination, noting that its absence necessitates costly custom software development for each dataset (Ruggles, 2018).

In addition to improving accessibility, metadata supports data quality and provenance tracking. This documentation of a dataset’s history and transformations ensures research

transparency and reproducibility (McQuilton et al., 2016). Such detailed curation allows users to assess data reliability and relevance, particularly in scientific research where integrity is paramount. By maintaining detailed metadata, organizations can ensure that their data assets remain valuable and trustworthy over time.

Finally, metadata contributes to the implementation of data governance and compliance with standards and regulations. Parmiggiani and Grisot identified three key elements of data curation – quality, protection, and relevancy filtering – as fundamental to effective data governance (Parmiggiani & Grisot, 2020). As data management standards evolve, compliance with frameworks such as FAIR becomes increasingly important. Metadata provides the necessary documentation to meet these standards, facilitating data sharing and reuse while maintaining legal and ethical considerations. By ensuring comprehensive and accurate metadata, organizations can enhance their data's long-term utility and return on investment (Harrow & Liener, 2021).

Funding agencies, universities, and government statistical agencies invest hundreds of millions of dollars in data infrastructure. Ruggles argues that the true value of these investments lies in the analytical power of the data, with long-term preservation being crucial for maintaining this value over time (Ruggles, 2018).

Importance of Standardized Metadata

Standardized metadata is a main component in RDM, and plays a key role in data usability, discoverability, interoperability, and long-term preservation. One of the primary benefits of standardized metadata is improved discoverability. When metadata follows consistent standards, it allows for effective cataloging and indexing in databases and repositories, making it easier for researchers and other stakeholders to search for and locate specific datasets (Duval

et al., 2002; Scherle, 2012). This improved discoverability ensures that valuable data can be found and utilized, increasing the impact of the research.

Additionally, standardized metadata increases interoperability, which is important for collaborative research and combining data from different sources and disciplines. When datasets are described using consistent metadata standards, they can be integrated and used together more efficiently, facilitating large-scale studies and meta-analyses (Wilkinson et al., 2016).

Standardized metadata also plays a role in maintaining data quality and consistency. By providing a consistent framework for describing data, standardized metadata helps ensure that data is accurately interpreted and used correctly in subsequent analyses (Batini et al., 2009). This consistency is a key for producing reliable research outcomes and maintaining the integrity of the data across different studies and applications.

Moreover, detailed and standardized metadata facilitates data reuse by allowing researchers to understand the context, provenance, and limitations of a dataset (Pampel & Dallmeier-Tiessen, 2014). This understanding is needed for reusing data appropriately, maximizing the value derived from existing data, and supporting the principles of open science. By making data more accessible and understandable, standardized metadata encourages more efficient and effective use of research resources (Azeroual et al., 2022).

Finally, standardized metadata supports long-term preservation efforts by ensuring that future users can understand and use the data, even if the original creators are no longer available to provide context. This is necessary for maintaining the accessibility and usability of data over time, as it ensures that data remains a valuable resource for future research (Conway, 2010). The use of metadata standards is therefore essential for preserving the longevity and utility of research data in the digital landscape (Garnett et al., 2017).

Differences in Metadata Requirements for Traditional vs. Open Datasets

The metadata requirements for traditional datasets and open datasets differ, reflecting the distinct objectives, accessibility, and usage context of each type. Traditional datasets, often managed within specific institutions or for specific projects, typically have more limited metadata requirements. The primary goal is to ensure that data can be effectively used and understood within the context it was collected. Metadata for traditional datasets generally focuses on context and provenance, access restrictions, and technical specifications. Open datasets, by contrast, are designed to be widely accessible and reusable by the broader research community and the public. As such, their meta requirements are more comprehensive and standardized to ensure usability across diverse context like discoverability, interoperability, licensing and use, quality and provenance, and user support and documentation (Duval et al., 2002; Michener, 2015; Pampel & Dallmeier-Tiessen, 2014; Scherle, 2012; Stodden et al., 2013; Wilkinson et al., 2016).

Economic Considerations and Costs in RDM

RDM involves several economic considerations that impact the efficiency, cost-effectiveness, and sustainability of data handling practices. Understanding these considerations is important for optimizing resources and ensuring that RDM practices contribute to the overall value and impact of scientific research.

One of the primary economic considerations in RDM is the cost of data storage and preservation. This includes expenses for physical storage infrastructure, cloud storage services, and the personnel required to maintain these systems. Long-term preservation also entails costs for migrating data to newer formats and technologies to prevent obsolescence. These costs must

be balanced against the benefits of ensuring data accessibility and usability over time (Beagrie, 2006; Keene, 2002).

RDM practices can influence the efficiency with which data is managed and shared. Ronald Coase's theory of transaction costs, originally formulated to explain the nature of firms and markets, can be applied to RDM to understand the costs associated with the transfer and sharing of data. Transaction costs in RDM include the time and resources spent on data curation activities such as cleaning, organizing, documenting, and ensuring compliance with standards and regulations (Coase, 1993). Reducing these costs through effective data management practices can lead to more effective and streamlined research processes (Arrow, 1969).

According to Wilkinson et al., data itself can be considered an economic asset, providing value through its potential to generate new knowledge, inform decision-making, and drive innovation. The economic value of data increases with its accessibility and reusability. RDM practices that enhance the FAIR Principles help maximize the economic return on investment in data collection and management by ensuring that data can be effectively used and reused by a broader audience (Wilkinson et al., 2016).

Another economic consideration is the securing of funding for RDM activities. Research institutions and funding agencies need to allocate resources not only for the initial stages of data collection and analysis, but also for the ongoing costs of data curation and preservation. Sustainable funding models are essential to ensure that RDM practices can be maintained over the long term. This includes exploring cost-sharing mechanisms, such as collaborative repositories and shared infrastructure, to distribute the financial burden (Mons et al., 2017).

An additional economic consideration in RDM is the opportunity costs which refer to the potential benefits relinquished when resources are allocated to certain data management

activities over others. For example, investing heavily in advanced storage solutions might limit the resources available for data analysis and interpretation. Balancing these opportunity costs involves strategic decision-making to ensure that investments in RDM yield the highest possible returns in terms of research impact and knowledge generation (Beagrie, 2006).

Compliance with data protection regulations and ethical standards also incurs economic costs. These include the costs of implementing data security measures, conducting regular audits, and ensuring that data management practices align with legal requirements. While these costs are necessary to protect data integrity and privacy, they must be managed efficiently to avoid undue financial burden on research projects (Borgman, 2013).

Economic Benefits of Open Data and Open Science

There are some specific economic benefits for the open science and open data initiatives. Piwowar and Vision believe that open science and open data practices enhance research efficiency and productivity by reducing duplication of efforts and enabling researchers to build on existing work. Access to a shared pool of data allows researchers to validate findings, conduct meta-analyses, and develop new hypotheses more rapidly (Piwowar & Vision, 2013). Tenopir et al., is of the same opinion as Piwowar and Vision, that researchers recognize the value of data sharing for validating and replicating scientific findings. By accessing shared datasets, scientists can independently verify results (Tenopir et al., 2011).

In their article “How Open Science Helps Researchers Succeed”, McKiernan et al. highlight how open data can foster innovation by providing raw material for developing new technologies, products, and services. Companies and startups can leverage publicly available data to create innovative solutions that address market needs and societal challenges (McKiernan et al., 2016). Open data can contribute to economic growth by enabling businesses and

governments to make data-driven decisions. Publicly available datasets can be used to optimize operations, improve service delivery, and identify new market opportunities, thus driving economic development (Ubaldi, 2013).

By sharing data openly, organizations can save costs associated with data collection and management. This collaborative approach reduces the need for redundant data collection efforts and spreads the cost across multiple entities, leading to overall cost efficiencies (Molloy, 2011). Tenopir et al. has the idea that data sharing among scientists improves research efficiency by being able to independently verify results and enables the concept of meta-analysis, which is the combination of multiple datasets. Meta-analysis aggregates findings from different studies to provide more robust and generalized conclusions, contributing to a deeper understanding of research questions (Tenopir et al., 2011).

Fecher and Friesike believe that open science practices increase transparency and accountability in research, which can enhance public trust in scientific findings and institutions. This trust is crucial for securing public and private funding for research initiatives, which in turn supports economic stability and growth in the research sector (Fecher & Friesike, 2014). Nosek et al. also believe open science practices improve transparency in the research process, making it easier to identify errors, verify results, and replicate studies. This transparency can lead to more efficient use of resources as researchers build on validated findings rather than duplicating efforts (Nosek et al., 2015).

Part of the economic benefits of open science is the facilitation of collaboration and networking among researchers, institutions, and industries across different regions and disciplines. According to Vicente-Saez and Marinez-Fuentes, this collaborative environment can lead to the cross-pollination of ideas, access to new funding sources, and the development of

inter-disciplinary projects, further driving economic benefits (Vicente-Saez & Martinez-Fuentes, 2018). Nosek et al. maintains that open science fosters collaboration by making data, methods, and findings openly available to the scientific community and others. They also believe that cross-disciplinary collaborations and innovative solutions can further drive economic benefits (Nosek et al., 2015).

Data Policies and Their Impact on Research

Data policies influence the research landscape by promoting open data practices, ensuring data protection, and increasing the transparency and reproducibility of scientific research. Key policies such as the NIH Data Management and Sharing Policy, Horizon Europe Open Science Policy, General Data Protection Regulation (GDPR), and NSF data management requirements play pivotal roles. By understanding and complying with these policies, researchers contribute to a more collaborative and innovative scientific community (Ayrís & Ignat, 2018; Burgess, 2020; National Institutes of Health, 2023; National Science Foundation, 2024).

The NIH has implemented policies to enhance data sharing and management in biomedical research. The NIH Data Management and Sharing Policy, effective January 25, 2023, requires researchers to submit data management and sharing plans with their grant applications. The goal of this policy is to maximize the availability of research data, promoting transparency, reproducibility, and data reuse. By mandating detailed data management plans, the policy ensures that data generated from NIH-funded research is accessible to the wider scientific community. For instance, NIH-funded researchers must describe how they will manage and share their data, including details on data types, related tools, standards, and how they will address privacy and confidentiality concerns. This encourages a culture of openness and collaboration in biomedical research (National Institutes of Health, 2023).

The European Union (EU) has several initiatives to promote open data and research transparency, including the Horizon 2020 and Horizon Europe programs. The EU Open Data Directive mandates that data generated by public sector bodies should be openly accessible and reusable. These policies promote research visibility and collaboration across member states, aiming to create a unified research area that facilitates innovation and addresses societal challenges through shared knowledge. For example, the Horizon Europe program requires that all research data be made available in a repository, with metadata provided to ensure discoverability. This supports the reuse of data across various disciplines, promoting interdisciplinary research and innovation (European Union, 2019).

The NSF requires all grant proposals to include a DMP detailing how data will be managed and shared. The NSF Public Access Plan mandates that publications and data resulting from NSF-funded research be made publicly accessible. These policies promote open access to research outputs, enhancing the transparency and reproducibility of scientific research. The requirements for DMPs ensure that data management practices are considered from the outset of the research project, improving data quality and availability. NSF-funded projects must submit a DMP outlining data types, standards, access policies, and plans for archiving and preservation, promoting a culture of responsible data stewardship and facilitating data sharing across the scientific community (*NSF - National Science Foundation*, 2024).

Search and Relevancy

In the late 1950s and early 1960s, the NSF sponsored a program to evaluate Information Retrieval (IR). In 1955, this effort led Cyril Cleverdon to create the Cranfield model, which requires a test collection of documents, a set of queries, and relevance judgements of relationships between the documents and queries. His first experiments were a database of

18,000 papers in the aeronautical engineering domain. The papers were indexed using four systems: The Universal Decimal Classification; an alphabetical subject index; a schedule of facet classification; and the Uniterm System of Coordinate Indexing (Kagolovsky & Moehr, 2003).

This experiment model was based on individual papers and did not represent a researcher's need to find relevant datasets. For each question, there was a single "core" relevant document, the one used to create the question. This experiment model exposed the need to provide a basic truth about the "relevance" of the search results. It was based on the researcher's own judgement; it featured no defined guidelines regarding what would make the returned results relevant to the query. The returned results were graded from one to five based on the user's thoughts. In this paper, the author defines relevancy as what the researcher perceives to be the most accurate result set returned, while using different search platforms and parameters.

Summary

This literature review chapter focused on the areas of open data, FAIR Principles, and data curation, all within the framework of RDM. The review aims to contribute to the broader discourse on effective data management practices, emphasizing the importance of making scientific research data more FAIR (Findable, Accessible, Interoperable, and Reusable). It also addresses the various issues in data curation and offers suggestions for improving these practices.

The chapter begins by outlining the increasing academic interest in RDM, demonstrated by a significant rise in published articles on the subject since 2012. This surge indicates a growing recognition of the importance of RDM in the research lifecycle, driven by policies from organizations such as the NIH and the NSF that mandate open research and RDM plans. Similar trends are observed in the number of publications on FAIR Principles and data curation, underscoring their relevance in contemporary research.

The chapter defines RDM and underscores its importance in the modern research landscape. RDM ensures the integrity, reproducibility, and usability of research data. Proper data management practices enable verification and building upon existing research, facilitating data sharing and collaboration among researchers. It also maximizes the value and longevity of research data, ensuring that data remains accessible and usable over time, even as technology evolves.

This chapter concludes by examining the economic benefits of open data and open science initiatives. These practices improve research efficiency and productivity, foster innovation, contribute to economic growth, and promote transparency and accountability in research. By understanding and complying with data policies from organizations such as NIH, the EU, GDPR, and NSF, researchers can contribute to a more collaborative and innovative scientific community.

CHAPTER 3

METHODOLOGY

Research Design

The foundation of this thesis is grounded in the principles of Information Retrieval (IR), a framework that guides the architecture of this research. IR systems are designed to fetch results in response to user queries, employing a variety of search techniques that differ across search engines. Among these techniques, keyword searching is fundamental, enabling researchers to enter specific terms or phrases, with the IR system scanning its database to identify relevant data. Additionally, Field Searching narrows this focus to document or dataset attributes, such as titles, authors, or subjects, enhancing precision. Moreover, the application of semantic search methodologies leverages Natural Language Processing to interpret the intent and contextual nuances of queries, transcending mere keyword matches.

This thesis encompasses two case studies, each utilizing the Search Efficiency Test methodology to evaluate the FAIR compliance of research datasets alongside the effectiveness of data curation processes. FAIR compliance, a concept championed by proponents of the FAIR Principles, signifies adherence to the principles of Findability, Accessibility, Interoperability, and Reusability. These case studies scrutinize the capabilities of the Data.gov and Google Dataset Search platforms, both noted for their IR search functionalities. Investigations were structured around defined search parameters – Title, Keywords, and Subject – applied consistently across both platforms to achieve predetermined search objectives.

The comparative analysis focused on these platforms, Data.gov and Google Dataset Search, stood at the core of this thesis' inquiry. The effectiveness of searches was quantified using a dataset collection $D = \{D1, D2, \dots, Dn\}$, with the goal to organize these datasets in

descending relevance to a specific query (q). This relevance was subjectively determined by the researcher's perception of how well the returned results matched their search intent. In quantifying relevance, datasets received ascending scores based on their order in the search results, reflecting their perceived relevance.

Acknowledging the subjective nature of relevance in IR, as highlighted by Cleverdon and Skopal, this thesis recognizes that relevance is influenced by the unique perspectives and information need of individual researchers (Škoda et al., 2020; Skopal et al., 2019). Thus, relevance is based on personal interpretations rather than an absolute standard.

Upon locating a dataset, further evaluation against the FAIR Principles' additional guidelines – accessibility, interoperability, and reusability – was conducted. This assessment included an extrapolation and validation of metadata, determining the dataset's curation success. These investigations aimed to address the research questions outlined below:

RQ1: What role does data curation play in the FAIR compliance of traditional research datasets and how does this role differ in the context of open datasets?

This question evaluates how data curation impacts the FAIR compliance of traditional versus open research datasets. It suggests that systematic organization, categorization, and metadata management significantly boost the FAIR compliance of traditional datasets. In contrast, open datasets depend on data curation tailored to standardize formats and metadata, catering to the diverse needs of the open data community, and ensuring broader accessibility and interoperability. This inquiry seeks to unravel the distinct roles these practices play in enhancing dataset FAIR compliance across different contexts (Palmer, 2009; Renear et al., 2010; Renear & Palmer, 2009).

RQ2: What is the role of metadata in enhancing the compliance of FAIR Principles in curated research datasets versus open datasets?

This investigation focuses on the fundamental role of metadata in reinforcing the FAIR Principles of datasets, examining both traditionally curated and openly available datasets. It suggests that metadata is crucial for making datasets more findable, accessible, interoperable, and reusable – essential aspects of FAIR compliance. Particularly in open datasets, the emphasis on standardized and exhaustive metadata practices is accentuated due to its extensive applicability across various user groups and fields. This thesis probes into how metadata's varying dynamics affect dataset FAIR compliance, contributing to a comprehensive understanding of metadata's integral role in RDM (Brickley et al., 2019; Devaraju & Berkovsky, 2018).

These inquiries were pursued through case studies on the Data.gov and Google Dataset Search platforms, chosen for their relevance and capability to procure targeted research datasets. These platforms provided a valuable lens for examining the interplay among FAIR compliance, data curation, and metadata, thereby enriching the understanding of their significance in research data management. The ensuing sections examine these platforms, offering insights into their functionalities.

Data.Gov

Data.gov is a comprehensive data catalog; it hosts a collection of more than 225,000 datasets meticulously gathered from various government agencies. The website covers a range of specific areas, including agriculture, climate, energy, local government, maritime, ocean, and health concerns pertaining to the elderly.

Data.gov utilizes schema.org and Data Catalog Vocabulary (DCAT) as its core metadata schema. Schema.org is an open, collaborative initiative designed to improve how structured data is represented on the internet. This structured data markup language enables annotation of web data and helps users understand web page content more accurately. DCAT is a standard developed to describe datasets and data services in a data catalog. It helps researchers find relevant datasets more easily because its standardized descriptions are more readily searchable and indexable. DCAT provides a common framework for data cataloging, enabling interoperability between different web-based data catalogs. It defines a structured format to describe datasets in a catalog, making it easier for these datasets to be discovered and integrated across different platforms. Alongside DCAT, the system incorporates the Resource Description framework, a family of World Wide Web Consortium specifications designed for data interchange on the web. Resource Description framework provides a framework to structure and link data in semantically meaningful ways, enabling more effective data discovery and reuse.

In addition to these foundational features, datasets available on Data.gov are subjected to a rigorous gathering process called “harvesting.” This involves collecting data from various departments and ensuring it adheres to specified DCAT and RD standards. Once harvested, the metadata of these datasets is transformed into a more accessible and widely used format, namely JSON (JavaScript Object Notation). This conversion facilitates downloading and manipulation of dataset metadata, thus allowing broader accessibility and application in various contexts. This comprehensive approach ensures that data on Data.gov is standardized, interoperable, and easily retrievable and usable by a broad range of researchers (Duke, 2022).

Google Dataset Search

Google Dataset Search made its debut in 2018. It provided the public with a specialized search engine tailored to discover datasets distributed across the internet. It does so by using two prominent and widely recognized vocabularies: Schema.org and the World Wide Web Consortium's Data Catalog. These vocabularies are integral to defining the underlying semantics of each dataset's individual webpages (Brickley et al., 2019).

Schema.org, a universally acknowledged standard, provides a structured framework to describe several types of data, facilitating comprehension and categorization of webpage content. It aids in the organization and presentation of dataset-related information, enhancing the search engine's ability to interpret and display relevant results accurately. Additionally, the World Wide Web Consortium's DCAT vocabulary complements schema.org by specifically focusing on dataset catalogs. It defines essential metadata properties and relationships related to datasets, enhancing the search engine's indexing and retrieval capabilities. This combination of schema.org and DCAT vocabularies ensures that Google Dataset Search can effectively navigate the intricate web of dataset information, delivering researchers comprehensive and meaningful search results (Halevy et al., 2016).

Google Dataset Search operates by systematically crawling metadata, assimilating and integrating information that is either growing or has been recently modified within web environments. This process is key to the aggregation of "web data," which denotes the collection of data directly sourced from the web's native environment. The continuous assimilation of this data is an important function; it elevates Google Dataset Search from basic search engine to robust data aggregator. In this capacity, it not only facilitates access to datasets, but actively gathers and presents metadata from diverse web sources (Sheridan et al., 2021).

Case Study Setup

In this thesis, research design focused on the methodology of case studies to examine the FAIR compliance and data curation practices across different data platforms. The cornerstone of this investigation was the application of Information Retrieval techniques to assess how effectively researchers could discover, access, evaluate interoperability, and determine the reusability of datasets. This inquiry was structured around a comparative analysis of two major platforms, Data.gov and Google Dataset Search, chosen for their IR capabilities and relevance to academic and open-source research communities.

Phase 1: Evaluating Dataset Findability

The initial phase of the case study methodology was dedicated to assessing the findability of datasets. This involved formulating specific queries (q) and then ranking the resulting dataset $D = \{D1, D2, \dots Dn\}$ based on their relevance to these queries. The determination of relevance was based on the researcher's subjective assessment of the alignment between the returned results and their search objectives. To quantify this relevance, dataset results were assigned scores in ascending order, according to their position in the search results, indicating their perceived relevance. This subjective assessment of relevance aimed to mirror the practical challenges researchers have when navigating IR platforms, reflecting the nuanced and often subjective nature of what constitutes relevant information. Relevance can be complex, and not clearly defined to researchers (Jansen & Pooch, 2001; Škoda et al., 2020; Skopal et al., 2019).

Phase 2: FAIR Compliance Assessment

Upon locating the datasets, the study proceeded to an evaluation of their FAIR compliance, using a detailed scorecard matrix to analyze each dataset's adherence to the FAIR Principles. The analysis covered:

- **Findability:** Assessment of unique identifiers, descriptive metadata, and registration in searchable resources.
- **Accessibility:** Evaluation of access protocols, data longevity, and access conditions.
- **Interoperability:** Consideration of data formats, use of standardized vocabularies, and metadata that facilitates integration.
- **Reusability:** Review of licensing information, provenance details, and adherence to community standards.

A closer look at specific data collected to evaluate FAIR compliance follows in Table 3.

Table 3

Dataset FAIR Compliance Evaluation

Guidelines	Guideline Descriptions
Findability	DOI or Globally Unique Identifiers (GUIDs) assigned to the dataset. Metadata that includes clear, accurate titles, descriptions, and keywords to improve discoverability. The registration or index information in a searchable resource.
Accessibility	Information on how to access the data, including credentials, URLs, or APIs. Information on the dataset's longevity, including where and how long the data will be stored. Clearly defined conditions under which the data can be accessed, including any restrictions or licenses.
Interoperability	Data should be in a format that is widely used and recognized by the community. Use of standardized vocabularies and ontologies for data description to ensure compatibility with other datasets. Metadata should include relationships to other data for integration and contextual understanding.

Guidelines	Guideline Descriptions
Reusability	Information on how the data can be used, including any restrictions or obligations. Detailed information about the data's origin and the methodology used to collect or generate it. Adherence to community standards and best practices for data documentation, quality, and security.

Note: Adapted from "*FAIR principles: Interpretations and implementation considerations*" by A. Jacobsen et al., 2020, Data Intelligence, 2(1-2), 10-29. In the public domain.

Table 4 identifies actual metadata fields that contain information to determine if the dataset has achieved FAIR compliance. Data.gov uses DCAT as the metadata format while Google Dataset Search uses Schema.org.

Table 4

FAIR Compliance Metadata Elements

FAIR Compliance Metadata	Data.gov DCAT Elements	Google Dataset Search schema.org
DOI or Globally Unique Identifiers (GUIDs)	'dct:identifier'	'identifier'
Titles	'dct:title'	'name'
Description	'dct:description'	'description'
Keyword(s)	'dct:keyword'	'keywords'
Registration/index information	'dct:isPartOf'	'includedInDataCatalog'
Credentials, URLs, or APIs	'dcat:Distribution:dcat:accessURL'	'distribution:contentUrl'
Where the data will be	n/a	n/a

FAIR Compliance Metadata	Data.gov DCAT Elements	Google Dataset Search schema.org
stored		
How long the data will be stored	n/a	n/a
Conditions, restrictions, licenses	‘dct:license’	‘license’
Data format(s)	‘dcat:Distribution:dct:mediaType’	‘distribution:encodingFormat’
Vocabularies and ontologies for data compatibility	‘dct:conformsTo’	Inferred from description
Relationships to other data	‘dct:relation’	‘isPartOf’, ‘hasPart’
How the data can be used	‘dct:license’	license or inferred from description
Data’s origin	‘dct:creator’	‘provider’
Methodology used to collect	‘dct:provenance’, ‘dct:source’	‘variableMeasured’ or ‘measurementTechnique’
Community standards and best practices for data documentation, quality, and security	‘dct:conformsTo’, Inferred from dct:description	‘schemaVersion’, inferred from description, ‘citation’

Note: Adapted from "Metadata resources and field mappings under the project open data metadata schema (DCAT-US schema v1.1)". (<https://resources.data.gov/resources/podm-field-mapping/#field-mappings>). In the public domain.

This evaluation offered a detailed perspective on each platform’s support for the FAIR Principles, underscoring the essential role of metadata in facilitating the FAIR compliance of datasets, making them findable, accessible, interoperable, and reusable.

Phase 3: Data Curation Analysis

The final phase focused on examining the data curation processes through the lens of metadata management. This involved a detailed analysis of:

- **Descriptive Metadata:** Evaluation of titles, descriptions, keywords, creator information, publication dates, and subject areas to gauge the dataset's clarity and searchability.
- **Administrative Metadata:** Review of rights, licensing, access restrictions, and preservation strategies to understand the dataset's governance and longevity.
- **Technical Metadata:** Assessment of file formats, data structures, versioning, and software requirements to determine technical accessibility and usability.
- **Provenance Metadata:** Examination of the dataset's origins, collection methods, and processing information to trace its lineage and integrity.
- **Use and Reuse Metadata:** Consideration of citation guidelines, related publications, and use cases to evaluate the dataset's impact and applicability.

A closer review of the metadata needed to analyze is presented in Table 5.

Table 5

Data Curation Evaluation

Metadata Type	Metadata Description
Descriptive Metadata	Title and Description provides a clear and concise summary of the data content. Keywords facilitate searchability and discovery. Creators/Authors provides identification for the person(s) responsible for the dataset. Publication Date provides when the data was published or released.

Metadata Type	Metadata Description
	Subject area is the field of study or domain where the data pertains to.
Administrative Metadata	<p>Rights and Licensing information about copyright, usage rights, and any licenses attached to the data.</p> <p>Access Restrictions are conditions or restrictions on accessing the data.</p> <p>Preservation Information details on data storage, backup procedures, and long-term preservation plans.</p>
Technical Metadata	<p>File format is the format in which the data is stored (e.g., CSV, JSON, XML).</p> <p>Data Structure is the description of the dataset structure, such as the schema or data model.</p> <p>Version information of the dataset if applicable.</p> <p>Software Requirements are software needed to access or use the data.</p>
Provenance Metadata	<p>Source where the data originated from, if not originally created by the authors.</p> <p>Collection methods detail how the data was collected, including instruments or techniques used.</p> <p>Processing information including transformations or processing the data underwent.</p>
Use and Reuse Metadata	<p>Citation instructions on how to cite the dataset.</p> <p>Related Publications include any publications that are based on or related to the dataset.</p> <p>Use Cases describe examples or case studies of how the data has been or can be used.</p>

Note. Adapted from: "*Understanding metadata*" by J. Riley, 2017. p. 7. In the public domain.

Analysis of the metadata elements for data curation expands the underlying metadata elements used in determining FAIR compliance of a dataset. A dataset may meet the FAIR compliance ideals but fail the data-curation process. A list of the metadata fields needed to analyze data curation follows in Table 6.

Table 6*Data Curation Metadata Fields*

Data Curation Metadata Elements	Data.gov DCAT Elements	Google Dataset Search schema.org
Title	'dct:title'	'name'
Keywords	'dcat:keyword'	'keywords'
Creators/Authors	'dct:creator'	'creator', 'author'
Publication Date	'dct:issued'	'datePublished'
Subject	'dct:subject'	'about'
Description	'dct:description'	'description'
Rights and Licensing	'dct:rights'	'license'
Access Restrictions	'dct:accessRights'	'conditionsOfAccess'
Preservation Information	'n/a'	'n/a'
File Format	'dcat:Distribution:dcat:MediaType'	'encodingFormat'
Data Structure	'dcat:Distribution:dct:format'	'variableMeasured'
Version Information	'dct:hasVersion'	'version'
Software Requirements Source	'n/a' 'dct:source'	'softwareRequirements' 'isBasedOn', 'sourceOrganization', inferred from 'description'
Collection Methods	Inferred from dct:Description	'measurementTechnique', Inferred from 'description'

Data Curation Metadata Elements	Data.gov DCAT Elements	Google Dataset Search schema.org
Processing Information	Inferred from dct:Description	‘measurementTechnique’, Inferred from ‘description’
Citation Instructions	‘dct:bibliographicCitation’	‘citation’
Related Publications	‘dct:references’, ‘dct:isReferencedBy’	‘citation’, ‘isReferencedBy’
Use Cases	‘n/a’, Inferred from ‘dct:Description’	‘exampleOfWork’, ‘hasPart’, Inferred from ‘description’

Note: Adapted from "*Metadata resources and field mappings under the project open data metadata schema (DCAT-US schema v1.1)*". (<https://resources.data.gov/resources/podm-field-mapping/#field-mappings>). In the public domain.

To facilitate the analysis, the metadata was collected and populated into a datasheet. This Metadata Collection Datasheet is comprised of two parts, the first section was used to capture the metadata elements to evaluate FAIR compliance as well as a recap of Phase 1 investigations. The second section was used to capture the metadata elements to evaluate the data curation elements. This is the form used to record the metadata values (Figure 7) as shown below.

Figure 7

FAIR Compliance Analysis - Metadata Collection Datasheet

		Metadata Collection Datasheet	
		Metadata FAIRness Elements	
Case Investigation #n	FAIRness based questions	Data.gov DCAT Elements	Google Dataset Search schema.org
Goal:	DOI or Globally Unique Identifiers (GUIDs) assigned to the dataset.	'dct:identifier'	'identifier'
Platform:	Titles to improve discoverability.	'dct:title'	'name'
Search Query:	Description to improve discoverability	'dct:description'	'description'
Parameter	Keyword(s) to improve discoverability	'dct:keyword'	'keywords'
Results:	Registration/Index information in a searchable resource	'dct:isPartOf'	'includedInDataCatalog'
Findable (Yes/No)	Information on how to access the data, including credentials, URLs, or APIs	'dcat:Distribution:dcat:accessURL'	'distribution:contentUrl'
Number of Results Returned	Information on where the data will be stored	n/a	n/a
Ranking of Dataset:	Information on how long the data will be stored	n/a	n/a
Accessible (Yes/No)	Under what conditions can the data be accessed, including restrictions or licenses	'dct:license'	'license'
Interoperable (Yes/No)	What data formats are available	'dcat:Distribution:dct:mediaType'	'distribution:encodingFormat'
Reusable (Yes/No)	What vocabularies and ontologies are used	'dct:conformsTo'	Inferred from description
	What relationships are included	'dct:relation'	'isPartOf', 'hasPart'
	How can the data be used, include restrictions or obligations	'dct:license'	license or inferred from description
	Where did the data originate from	'dct:creator'	'provider'
	What methodology was used to collect the data	'dct:provenance', 'dct:source'	'variableMeasured' or 'measurementTechnique'
	Adherence to community standards and best practices. Is Documentation available	'dct:conformsTo', Inferred from dct:description	'schemaVersion', inferred from description, 'citation'

Note: Adapted from "Metadata resources and field mappings under the project open data metadata schema (DCAT-US schema v1.1)". (<https://resources.data.gov/resources/podm-field-mapping/#field-mappings>)

Figure 8 is the form used to collect data curation metadata.

Figure 8*Data Curation Analysis - Metadata Collection Datasheet Analysis*

Case Investigation #n	Data Curation based Questions	Data Curation Metadata Elements	
		Data.gov DCAT Elements	Google Dataset Search schema.org
Goal:	Title provides a clear concise summary of the data contents	'dct:title'	'name'
Platform:	Description provides a clear concise summary of the data contents	'dcat:keyword'	'keywords'
Search Query:	Keywords facilitate search and discovery	'dct:creator'	'creator', 'author'
Parameter	Creators/Authors provide identification for the person(s) responsible for the dataset	'dct:issued'	'datePublished'
	Publication date provides when the data released or published	'dct:subject'	'about'
	Subject area is the field of study or domain of the data	'dct:description'	'description'
	Rights and Licensing info about copyright, usage rights, or other licenses	'dct:rights'	'license'
	Access Restrictions or conditions on accessing data	'dct:accessRights'	'conditionsOfAccess'
	Preservation information for data storage, backup procedures, long-term preservation plans	'n/a'	'n/a'
	File formats the data is stored in	'dcat:Distribution:dcat:MediaType'	'encodingFormat'
	Data Structure description, schema or model	'dcat:Distribution:dct:format'	'variableMeasured'
	Version information	'dct:hasVersion'	'version'
	Software requirements needed to access the data	'n/a'	'softwareRequirements'
	Source where the data originated from	'dct:source'	'isBasedOn', 'sourceOrganization', inferred from 'description'
	Collection methods, including how it was collected, including instruments or techniques	Inferred from dct:Description	'measurementTechnique', Inferred from 'description'
	Transformations or processing the data underwent	Inferred from dct:Description	'measurementTechnique', Inferred from 'description'
	Citation instructions	'dct:bibliographicCitation'	'citation'
	Related publications based or related to dataset	'dct:references', 'dct:isReferencedBy'	'citation', 'isReferencedBy'
	Use Case examples or case studies on how the data has been or can be used	'n/a', Inferred from 'dct:Description'	'exampleOfWork', 'hasPart', Inferred from 'description'

Note: Adapted from "Metadata resources and field mappings under the project open data metadata schema (DCAT-US schema v1.1)". (<https://resources.data.gov/resources/podm-field-mapping/#field-mappings>)

As noted in Phase 3 – Data Curation Analysis requires many of the same metadata elements identified in Phase 2 of the case study investigations. A list of these overlapped elements is in Table 7 displayed below.

Table 7*Metadata FAIR Principles and Data Curation Overlapped Elements*

Description of Metadata Elements	Data.gov DCAT Elements	Google Dataset Search schema.org
Title provides a clear concise summary of the data contents	'dct:title'	'name'

Description of Metadata Elements	Data.gov DCAT Elements	Google Dataset Search schema.org
Description provides a clear concise summary of the data contents	'dcat:keyword'	'keywords'
Subject area is the field of study or domain of the data	'dct:description'	'description'
What data formats are available	'dcat:Distribution: dct:mediaType'	'distribution:encodingFormat'

This granular focus on metadata illuminated the data curation practices of each platform and underscored overlapping elements between FAIR Principles and effective data curation. By highlighting the role of well-managed metadata, this case study methodology offered insights into the mechanisms that underpin dataset FAIR compliance and the efficacy of data curation efforts to enhance the value and utility of research data.

Title Search

The first case investigation for Phase 1 – “Evaluating Dataset Findability,” used selected “Title” as the search parameter. Using “Title” is a crucial feature that enhances the efficiency of platforms, facilitating researchers’ efforts to find specific datasets quickly and accurately. Searching by “Title” also allows researchers to find datasets by their specific names. It is best used when researchers know the exact dataset they seek, or at least the precise terms used in the dataset’s title. The “Title” search is a quick and straightforward way to access information researchers need without having to sift through unrelated data. With large repositories—and one of the two platforms featured voluminous data—it reduces time spent browsing through extensive dataset lists. Searching by title can help researchers determine if a dataset already

exists, avoiding duplication of efforts in data collection or analysis. The Title Search was performed on the platform Data.gov and Google Dataset Search.

To continue with Phase 2 – FAIR Compliance Assessment, metadata was captured from the Data.gov and Google Dataset Search platforms as well as from the dataset itself. To evaluate FAIR compliance the metadata elements from Table 4 were extracted and populated into the Metadata Collection Datasheet. For Phase 3 - Data Curation Analysis, metadata elements from Table 6 were extracted from the dataset and populated into the Metadata Collection Datasheet.

Keyword Search

The second case investigation used “keyword(s)” as the search parameter for the Evaluating Dataset Findability Phase 1. Unlike “Title” searches, which require specific knowledge of a dataset’s name, keyword searchers allow researchers to locate datasets based on a wide range of terms related to content, theme, or subject matter. This makes data more accessible to researchers who may not know the exact titles of the datasets they need. Keyword searches help researchers discover datasets related to their area of interest, but perhaps not immediately obvious from the dataset title alone. They are particularly useful for research and analysis, as they expose researchers to a wider spectrum of relevant data. Keyword search facilitates exploratory data analysis by allowing researchers to start with a broad topic and refine their search as they learn more about the available data, which makes exploration both educational and efficient. Overall “Keyword” search enhances the discoverability of datasets, supports diverse researcher needs, and facilitates effective data exploration and utilization. The case investigation was conducted by selecting specific keywords and entering them into two platforms: Data.gov and Google Dataset Search.

Proceeding to Phase 2 – FAIR Compliance Assessment, metadata was gathered from both Data.gov and Google Dataset Search platforms, in addition to the dataset itself. To assess FAIR compliance, the extracted metadata elements outlined in Table 4 were entered into the Metadata Collection Datasheet. In Phase 3 – Data Curation Analysis, metadata elements are included in Table 6 from the dataset and in the Metadata Collection Datasheet.

Subject Search

The third case investigation focused on “Subject” searching. Subject searches are similar in many ways to keyword searches, but key differences exist. For instance, “Subject” searches are precise and structured; keyword searches tend to be more flexible and researcher friendly. Whereas Subject searches are ideal when researchers know the exact topic they are pursuing, Keyword searches are more suitable for broader explorations and when a specific subject heading is unknown. A Subject search usually indicates a predefined set of topics or categories, or inclusion in a controlled vocabulary or classification system. A Keyword search looks for specified words or phrases anywhere in content text or metadata. This could include titles, descriptions, full text, etc.

Subject search headings are standardized, and therefore can provide more consistent results. This consistency can deliver a more precise search result because results are categorized for a specific Subject heading, ensuring that all results are closely related to the topic. The case investigation was performed by entering a Subject into Data.gov and Google Dataset Search. Both platforms were evaluated as to whether the selected Subject was returned, and their rankings.

Moving forward to Phase 2 – FAIR Compliance Assessment, involved collecting metadata from the Data.gov and Google Dataset Search platforms, along with the dataset itself.

For evaluating FAIR compliance, the extracted metadata elements specified in Table 4 were recorded into the Metadata Collection Datasheet. Similarly, for Phase 3 – Data Curation Analysis, the metadata elements listed in Table 6 from the dataset were entered into the Metadata Collection Datasheet.

Case Study Recap

Three distinct case investigations were conducted across two search platforms to evaluate their performance using various search methods. Case investigation 1 used “Title” as the search parameter in the Data.gov and Google Dataset Platform. Case investigation 2 used a “Keyword” search as the parameter also using Data.gov and Google Dataset platforms. In the final case investigation—Case Investigation 3— “Subject” was input as the search parameter in both platforms. The objective was to determine the efficacy of each platform’s ability to return relevant search results and to analyze the rankings of these results.

Once the datasets were retrieved, further metadata analysis was performed by capturing and extracting various metadata elements related to both the FAIR Principles and data curation and input into a spreadsheet for further analysis. This comprehensive approach delivered a comparative understanding as to how each platform managed different search methodologies.

Validation

A thorough verification process was implemented to validate the accuracy and relevance of the information obtained in the case studies. The process was essential to ensure the reliability of the findings and to confidently draw meaningful conclusions. The validation process involved several key steps:

Link Verification: For each dataset identified in the search results, the corresponding link was selected and followed. This step helped ascertain that the link accurately directed the

researcher to the intended dataset. It served as a primary check for the functionality and correctness of the search results provided by the platforms.

Visual Inspection of Datasets: Once directed to the dataset via its link, the researcher conducted a visual inspection of the dataset. It was reviewed to verify that its contents matched expectations set by the search query and the initial search result description. Visual inspection enabled a qualitative assessment of the dataset's relevancy and suitability for the research purpose.

Metadata Comparison: The comparison of metadata was an important aspect of validation. The researcher examined whether metadata displayed in the search results aligned with the actual metadata associated with the dataset at its source. This step confirmed the accuracy of information provided in the search results. It also ensured that the metadata accurately represented the dataset's content.

Consistency Across Search Methods: This verification process was applied consistently across all three search methods (Title, Keywords, Subject) on both Data.gov and Google Dataset Search. Validation approach consistency helped ensure that the case investigation findings were robust and applicable across different search methodologies and platforms.

Documenting Inconsistencies: Throughout validation, the researcher documented inconsistencies or issues, such as broken links, mismatched metadata, or irrelevant dataset content. This process identified the search platforms' limitations and challenges and facilitated a broader evaluation of their effectiveness.

Table 8*Validation Process Descriptions*

Validation Process Type	Validation Process Description
Link Verification	Select the link and verify the result dataset is the correct one
Visual Inspection	Verify visually that the dataset contents match expectations set by the search query
Metadata Comparison	Verify the metadata matches the source search results
Consistency	Verify that the verification process was applied across all search methods
Documentation	Document inconsistencies or issues

By implementing these validation steps, the case investigations confirmed that datasets identified as relevant were indeed pertinent and accurately described. The thorough validation process established the case study findings' credibility and reliability.

Documentation

The researcher used several forms as part of the case investigations. It provided a vehicle for results analysis, and captured results returned from each case investigation. The information includes the platform being searched, the type of search, the search parameter used, and a series of yes or no questions regarding the FAIR Principles. Information gathered from results was entered into forms based on one of three phases, which included each case investigation's goal (Figure 9).

Figure 9*Evaluating Dataset Findability - Metadata Collection Datasheet*

Case Investigation #n	Results
Goal: Platform: Search Query: Parameter: Title Results: Findable (Yes/No) Number of Results Returned: Ranking of Dataset: Accessible (Yes/No) Interoperable (Yes/No) Reuseable (Yes/No)	

During Phase 2 and Phase 3 of the case study investigations, additional sections of the Metadata Collection Datasheet were populated with captured metadata, based on the FAIR compliance questions (Figure 10) and data curation questions (Figure 11).

Figure 10*FAIR Compliance Assessment – Metadata Collection Datasheet*

FAIRness based questions	Metadata FAIRness Elements	
	Data.gov DCAT Elements	Google Dataset Search schema.org
DOI or Globally Unique Identifiers (GUIDs) assigned to the dataset.		
Titles to improve discoverability.		
Description to improve discoverability		
Keyword(s) to improve discoverability		
Registration/Index information in a searchable resource		
Information on how to access the data, including credentials, URLs, or APIs		
Information on where the data will be stored		
Information on how long the data will be stored		
Under what conditions can the data be accessed, including restrictions or licenses		
What data formats are available		
What vocabularies and ontologies are used		
What relationships are included		
How can the data be used, include restrictions or obligations		
Where did the data originate from		
What methodology was used to collect the data		
Adherence to community standards and best practices. Is Documentation available		

Figure 11

Data Curation Analysis - Metadata Collection Datasheet

Data Curation based Questions	Data Curation Metadata Elements		
	Data.gov DCAT Elements	Google Dataset schema.org	Search
Title provides a clear concise summary of the data contents			
Description provides a clear concise summary of the data contents			
Keywords facilitate search and discovery			
Creators/Authors provide identification for the person(s) responsible for the dataset			
Publication date provides when the data released or published			
Subject area is the field of study or domain of the data			
Rights and Licensing info about copyright, usage rights, or other licenses			
Access Restrictions or conditions on accessing data			
Preservation information for data storage, backup procedures, long-term preservation plans			
File formats the data is stored in			
Data Structure description, schema or model			
Version information			
Software requirements needed to access the data			
Source where the data originated from			
Collection methods, including how it was collected, including instruments or techniques			
Transformations or processing the data underwent			
Citation instructions			
Related publications based or related to dataset			
Use Case examples or case studies on how the data has been or can be used			

Summary

This thesis uses the principles of Information Retrieval (IR). IR are tools for retrieving relevant results based on user queries, employing various techniques such as keyword searching, field searching, and top searching. These methods allow for both broad and precise data retrieval. This research includes several case investigations that employed the Search Efficiency Test methodology to access the FAIR compliance of research datasets and evaluate the data curation process via metadata. The two platforms that case investigations were focused on were Data.gov and Google Dataset Search. Using search parameters – Title, Keywords, and Subject – to achieve defined search objectives. The comparative analysis of these platforms revealed insights into the effectiveness of searches, with relevance scored based on the perceived alignment of results with the search intent. The subjective nature of relevance in IR was acknowledged, emphasizing the influence of individual researchers’ perspectives on the outcome of searches. Further, the thesis evaluated datasets against additional FAIR guidelines by analyzing their metadata and curation success. In the next section of this thesis, the findings of these case investigations are discussed.

CHAPTER 4

RESULTS

Findings Overview

The central focus of this thesis is data curation and FAIR compliance of research datasets, in both traditional research datasets and research open datasets. Navigating the FAIR Principles has served as a framework for establishing FAIR compliance. FAIR Principles calls for data to be Findable, Accessible, Interoperable, and Reusable. If the dataset meets the FAIR criteria, it is considered to embody FAIR compliance.

The Findability guideline, the ‘F’ in the FAIR Principles, calls for data to feature an internationally distinct and enduring identifier, as well as comprehensive linked descriptive information. Moreover, these components should be duly registered or cataloged within a searchable repository. The scope of findability for datasets is broad. Options include datasets coupled with their corresponding metadata, and direct associations between dataset components and identifiers. Alternatively, findability could involve only the dataset or metadata alone, without direct linkage to specific dataset components (Turner et al., 2022; Wilkinson et al., 2016).

The Accessibility guideline, the ‘A’ in the FAIR Principles, emphasizes the necessity for data to be readily retrievable upon discovery. According to Wilkinson et al., for data to be deemed Accessible, it must meet specific criteria. First and foremost, both the data and its accompanying metadata should be obtainable via an identifier through a standard communications protocol that is open and free, as well as universally implementable. Such protocols, which might include HTTP, HTTPS, FTP, and SFTP, should ideally support authentication and authorization mechanisms to ensure controlled access, particularly for

sensitive data (Wilkinson et al., 2016). Additionally, another Accessibility requirement Wilkinson et al. highlighted is that metadata must remain accessible, even in instances where the corresponding data has been deleted or is no longer available. This persistence ensures that descriptive details about the dataset are preserved, thereby supporting ongoing data discovery and comprehension efforts, even without the original dataset's availability (Wilkinson et al., 2016).

Interoperability, the 'I' in FAIR, focuses on the ability of data and systems to work together within and across organizational boundaries to meet the needs of various stakeholders. This involves the seamless exchange and use of data across different platforms, applications, and environments. Interoperability requires data and metadata to be represented in a standardized, shareable format that is understandable and usable by different systems. This is achieved using common standards, vocabulary, and frameworks that ensure data from diverse sources can be integrated and analyzed together. Examples of common standard data formats include CSV, JSON, and XML (Wilkinson et al., 2016, 2017).

The 'R' in FAIR stands for Reusability. It is one of the core concepts that drives the open science movement. This guideline reflects the need for data to be effectively used and reused over time, both within and beyond the original context of its collection. Reusability is important for advancing scientific discovery and maximizing the return on the investment made in collecting and curating data. The criteria for meeting Reusability are i.) Data should have detailed metadata and documentation that provides clear, accessible, and comprehensive information about the data, including its context, quality, and condition, and any other information necessary for its reuse, ii.) Data and metadata should adhere to community standards regarding format, vocabulary, and protocols, iii.) Licenses for use, reuse, and sharing should be

clear and explicitly defined, iv.) Information about the origin, methodology, and processing of data should be documented to provide provenance for it (Jacobsen et al., 2020; Wilkinson et al., 2016).

This thesis investigated two research questions relating to the role of data curation and the significance of metadata in the context of FAIR compliance in research datasets:

RQ1: What role does data curation play in the FAIR compliance of traditional research datasets, and how does this role differ in the context of open datasets?

Assumption: Data curation was anticipated to enhance the FAIR compliance of traditional research datasets significantly through structured organization, categorization, and metadata management. Conversely, an approach focusing on format standardization and metadata was likely more effective for open datasets. That method would address their need for wider accessibility and interoperability.

Expected Outcome: The study was designed to demonstrate that meticulous data curation significantly improved FAIR compliance in traditional research datasets. In open datasets, an emphasis on standardization metadata and formats was expected to enhance FAIR compliance, thus aligning with FAIR principles.

RQ2: What is the role of metadata in enhancing the FAIR compliance of curated research datasets versus open datasets?

Assumption: The thesis suggests that metadata is essential to the FAIR compliance of both curated traditional research datasets and open datasets, with an emphasis on metadata standardization in open datasets.

Expected Outcome: The research sought to confirm that metadata significantly influenced FAIR compliance of both traditional and open research datasets. It was expected that

metadata's comprehensive, structured nature would help facilitate FAIR compliance of datasets in curated traditional research settings. In open datasets, the thesis aimed to highlight how standardized metadata practices enhance accessibility and interoperability, underscoring their prominence compared to traditional datasets.

Regarding RQ1—the role data curation plays in the FAIR compliance of traditional research datasets and its difference in the context of open research datasets—a multifaceted approach seemed optimal. This thesis theorized that data curation was integral in the Findability, Accessibility, Interoperability, and Reusability of traditional research datasets. This result would be achieved through systematic organization, categorization, and effective metadata management. Such practices in traditional datasets ensure that data is easily discoverable and accessible.

By contrast, data curation remained a critical element of open datasets, but required a different approach: a focus shifted toward data format and metadata standardization (Palmer, 2009; B. Zhang et al., 2016). This adjustment was considered essential to ensure broader accessibility and interoperability of research open datasets. The rationale behind the expectations of this case study was that open datasets cater to a more diverse community, encompassing various disciplines and user needs. Therefore, standardizing how data was curated, particularly in terms of formats and metadata, should have been more beneficial in addressing wide-ranging requirements of the open-data community. This approach aligned with existing literature and research in the field—as highlighted by (Palmer, 2009; Renear et al., 2010; Renear & Palmer, 2009)—which emphasized the importance of tailored data curation practices to meet specific needs of different data types and user communities.

RQ2, which considered the role of metadata in both curated research datasets and open datasets, asked about a different, but crucial, metadata role in these contexts. The assumption suggested that metadata is a vital component that ensures both the findability, accessibility, interoperability, and reusability of data, regardless of whether it is part of a curated traditional dataset or a research open dataset.

However, there was a subtle difference as to how metadata influenced the two specific datasets discussed. Based on research and case investigation results, metadata is more prominent in research open datasets. This distinction was attributed to the need for more standardized and comprehensive metadata practices in open datasets. Such practices were considered to enable effective data integration and reuse across platforms and disciplines, thereby enhancing the utility of open datasets.

Conversely, while metadata is still important with curated traditional research datasets, research and case investigations show that they feature less flexibility in how metadata is structured and applied. This rigidity may have been due to the more controlled environment in which curated traditional research datasets were managed. In such environments, there may be more room to tailor metadata to specific organizational or contextual needs.

This case study expectation emanated from the understanding that metadata is a fundamental tool to organize and access data. This concept was supported by field research, including the works of Brickley et al. (Brickley et al., 2019; Devaraju & Berkovsky, 2018). The studies highlighted varying requirements and applications of metadata in different management contexts, underlining its ability to FAIR compliance.

The thesis author conducted a Search Efficiency Test to answer both research questions. The expected outcome was that Data.gov would provide more datasets relevant to the search

parameters. It would also illustrate that standardized metadata is more effective. Additionally, the outcome was expected to show that highly curated datasets, such as those in Data.gov, would provide more precise dataset FAIR compliance.

The author conducted six case investigations, using three different searches over two platforms. One platform was Google Dataset Search, a data aggregator whose datasets are presented as links to users. They are also crawled and indexed. The original dataset owners provided and updated metadata. The second platform was Data.gov, a data catalog, which contains open datasets curated by the U.S. government. Dataset links are created and maintained by various U.S. government departments.

A quick view of the results of the case investigations dataset findability follows (Table 9).

Table 9

Evaluating Dataset Findability Case Investigations Results Overview

Platform	Data.gov	Google Dataset Search
Title Search	Case Investigation 1 # of Results Returned: 171 Specific Dataset found: Yes Ranking of found Dataset: 1	Case Investigation 1 # of Results Returned: 100+ Specific Dataset found: Yes Ranking of found Dataset: 1
Keyword Search	Case Investigation 2 # of Results Returned: 13,326 Specific Dataset found: Yes Ranking of found Dataset: 2	Case Investigation 2 # of Results Returned: 99 Specific Dataset found: Yes Ranking of found Dataset: 5
Subject Search	Case Investigation 3 # of Results Returned: 789 Specific Dataset found: Yes Ranking of found Dataset: 1	Case Investigation 3 # of Results Returned: 100+ Specific Dataset found: Yes Ranking of found Dataset 1

According to these results, in Phase 1 - Evaluating Dataset Findability, all case investigations showed that the specific targeted dataset was found on both platforms. They also demonstrated several hundred results returned. In each case, the ranking of the targeted dataset was within either the first one or within the first five in reference to “Keyword” search results. Details of each case investigation and the results of Phase 2 – FAIR Compliance Assessment, and Phase 3 - Data Curation Analysis, were analyzed further in the following sections.

Case Investigation 1 – Title Search

A specific dataset title, “Mental Health Care in the Last 4 Weeks,” was selected as the target from Data.gov. As part of ongoing Covid-19 programs, government agencies were trying to find data about the pandemic’s social and economic impact on American households. The data they sought included the state of mental health care. As such, starting September 23, 2020, the government created a recurring four-week dataset based on an Internet questionnaire distributed to U.S. households by the U.S. Census Bureau. It was published by the Centers for Disease Control and Prevention and submitted into the Data.gov repository. The last time this dataset was updated was April 15, 2023. In each platform, the “Search” parameter was entered, and results were returned. A review of Phase 1 - Evaluating Dataset Findability results of Case Investigation 1 – Title Search is presented in Table 10.

Table 10

Case Investigation 1 – Title Search Results

Case 1 Investigation – Title Search	Data.gov	Google Dataset Search
Goal:	Find the most recent US Government data on the social and economic impacts of Covid-19 on mental health focusing on	Find the most recent US Government data on the social and economic impacts of Covid-19 on mental health focusing on

Case 1 Investigation – Title Search	Data.gov	Google Dataset Search
	the HHS department data catalog	the HHS department data catalog
Platform:	Data.gov	Google Dataset Search
Search Query:	Mental Health Care in the Last 4 Weeks	Mental Health Care in the Last 4 Weeks
Parameter:	Title	Title
Results: Findable (Yes/No)	Yes	Yes
Number of Results Returned:	171	100+
Ranking of Dataset:	1	1
Accessible (Yes/No)	Yes – Based on Platform Accessibility	Yes – Based on Platform Accessibility
Interoperable (Yes/No)	Yes – Based on Platform Interoperability	Yes – Based on Platform Interoperability
Reusable (Yes/No)	Yes – Based on Platform Criteria	Yes – Based on Platform Criteria

Results of Case Investigation 1 – Title Search, show that both platforms returned the specific title “Mental Health Care in the Last 4 Weeks.” They also show that Data.gov returned 171 additional records and Google Dataset Search returned more than 100. Each ranking for the dataset returned is 1, so this title “Mental Health Care in the Last 4 Weeks” was first in the results list.

The FAIR compliance criteria have a visual component highlighted by reviewing the Accessible, Interoperable, and Reusable criteria. Using platform standards as they exist, researchers can see that they have access to the dataset via both Data.gov and Google Dataset Search. They can also determine interoperability standards exist by noticing they can select different file formats for dataset downloads. Researchers can also assume that since they can find the dataset and access it, they can reuse it, based on the perception provided by the platforms.

In Phase 2 – FAIR Compliance Assessment, the actual metadata from each platform is gathered. Table 11 is a worksheet containing numerous detailed questions regarding dataset outcomes used in Case Investigation 1. Each worksheet question is associated with a FAIR Principle. For further detailed metadata elements, Google Dataset Search does not provide all metadata in any visible or downloadable methods. Metadata provided to Google Dataset Search from Data.gov may be available via web-scraping technology. However, from a researcher's perspective, that information is contained in a black box; only the visible elements for this thesis are included in this analysis when reviewing Google Dataset Search results. Conversely, Data.gov features a link on the search results web page where metadata can be opened, inspected, and downloaded.

Table 11*Case Investigation 1 – Title Search - FAIR Compliance Assessment*

FAIR Principle	FAIR Compliance Questions	Data.gov DCAT	Google Dataset Search schema.org
Findability	DOI or Globally Unique Identifiers (GUIDs) assigned to the dataset?	Yes	No
Findability	Titles to improve discoverability?	Yes	Yes
Findability	Description to improve discoverability?	Yes	Yes
Findability	Keyword(s) to improve discoverability?	Yes	No
Accessibility	Registration/Index information in a searchable resource?	Yes	No
Accessibility	Information on how to access the data, including credentials URLs, or APIs?	Yes	Yes
Accessibility	Information on where the data will be stored?	No	No
Accessibility	Information on how long the data will be stored?	No	No
Accessibility	Under what conditions can the data be accessed, including restrictions or licenses listed?	Yes	No
Interoperability	What data formats available listed?	Yes	Yes
Interoperability	What vocabularies or ontologies listed?	Yes	No

FAIR Principle	FAIR Compliance Questions	Data.gov DCAT	Google Dataset Search schema.org
Interoperability	What relationships are included listed?	Yes	No
Reusability	How can the data be used, including restrictions or obligations listed?	Yes	No
Reusability	Where did the data originate listed?	Yes	No
Reusability	What methodology was used to collect the data listed?	Yes	No
Reusability	Adherence to community standards and best practices. Is documentation available?	No	No

Among the 16 questions designed to evaluate adherence to FAIR compliance criteria, Data.gov provided responses to 14, while Google Dataset Search addressed only four. Specifically, for the Findability aspect of the FAIR Principles, Data.gov supplied an identifier, title, description, and keywords for the dataset “Mental Health Care in the Last 4 Weeks,” demonstrating compliance with the Findability standards. In contrast, an examination of the same dataset through Google Dataset Search revealed answers to two questions – title and description – without any visible identifier or keywords listed.

The Accessibility component of the FAIR compliance guidelines was evaluated through five specific questions, with Data.gov fulfilling three by offering details on indexing, URL location, and licensing. However, it lacked information regarding the storage location and duration of data storage. Conversely, Google Dataset Search met one of these five criteria by

indicating the dataset’s URL location, without providing details on indexing, storage, or licensing conditions.

In the evaluation concerning the Interoperability FAIR Principle, which included three questions, Data.gov successfully addressed all, covering data formations, vocabularies, and the dataset’s relationships with other data. Google Dataset Search, however, only managed to respond to the questions about data formats, leaving vocabularies and relationships unaddressed.

Regarding the Reusability FAIR Principle, which encompassed four questions, Data.gov provided answers to three, but failed to offer metadata related to adherence to community standards or documentation of best practices. However, Google Dataset Search did not address any of the four questions related to reusability, lacking information on data restrictions, data origin, collection methodologies, and adherence to community standards and practices.

To continue the analysis for Case Investigation 1, the curation data for Phase 3 - Data Curation Analysis, was gathered and populated into the Data Curation Metadata Element worksheet. Questions were based on the needs for data curation. Each question denotes the FAIR Principle with which it is associated. Answers came from results returned in Case Investigation 1 and are based on available metadata. Table 12 helps analyze data curation.

Table 12

Case Investigation 1 – Title Search – Data Curation Analysis

FAIR Principle	Data Curation Questions	Data.gov DCAT	Google Dataset Search schema.org
Findability	Title provides a clear concise summary of the data contents?	Yes	Yes
Findability	Description provides a clear concise summary of the data contents?	Yes	Yes

FAIR Principle	Data Curation Questions	Data.gov DCAT	Google Dataset Search schema.org
Findability	Keywords facilitate search and discovery?	Yes	No
Findability	Creators/Authors provide identification for the person(s) responsible for the data?	Yes	Yes
Findability	Updated Metadata updated date?	Yes	Yes
Findability	Publication date provides when the data released or published?	Yes	No
Findability	Subject area is the field of study or domain of the data?	Yes	No
Accessibility	Rights and Licensing info about copyright, usage rights, or other licenses?	Yes	No
Accessibility	Access Restrictions or conditions on accessing data?	Yes	No
Accessibility	Preservation information for data storage, backup procedures, long-term preservation plans?	No	No
Interoperability	File formats the data is stored in?	Yes	Yes
Interoperability	Data Structure description, schema or model?	No	No
Interoperability	Version information?	No	No
Interoperability	Software requirements needed to access the data?	No	No
Reusability	Source where the data originated from?	Yes	No

FAIR Principle	Data Curation Questions	Data.gov DCAT	Google Dataset Search schema.org
Reusability	Collection methods, including how it was collected, including instruments or techniques?	Yes	Yes
Reusability	Transformations or processing the data underwent?	Yes	Yes
Reusability	Citation instructions?	No	No
Reusability	Related publications based or related to dataset?	No	No
Reusability	Use Case example or case studies on how the data has been or can be used?	No	No
Reusability	Publisher?	Yes	Yes

Of 21 data curation inquiries, Data.gov's metadata successfully addressed 14, as evidenced by Table 12. Google Dataset Search furnished responses to only eight of the 21 questions. Examining their conformity with the FAIR compliance standards, particularly the Findability aspect, which constituted seven questions, both Data.gov and Google Dataset Search supplied metadata for titles, descriptions, creators, and updated metadata dates. However, Google Dataset Search did not offer metadata on keywords, publication dates, or subject areas.

For the three questions pertaining to the Accessibility principle of FAIR compliance, Data.gov responded to two, providing metadata regarding licensing and data access restrictions, but lacking information on data-preservation methods. Google Dataset Search, however, provided metadata responses for no Accessibility-related questions.

The analysis worksheet for the Interoperability guideline included four questions. Both Data.gov and Google Dataset Search addressed just one related to the availability of file formats. The remaining queries about data structure or model, version information, and software requirements went unanswered by both platforms.

The Reusability section posed seven questions. Data.gov's answered four, detailing the dataset's source, collection methods, any processing it underwent, and the publisher. It did not include citation guidelines, related publications, or use cases exemplifying how the data might be used. On the other hand, Google Dataset Search supplied metadata for three of the seven Reusability questions but did not include the dataset's origin source.

Case Investigation 1 – Title Search - Validation

In Case Investigation 1 – Title Search, the first platform the “Title” search was performed on was Data.gov. From here, 171 datasets were returned, with the targeted dataset returned as the first. To determine if the link provided was valid and did indeed point to the targeted search dataset, the researcher navigated to Data.gov and entered the title into the search box. Below, Figure 12 shows the results of the navigation and search results.

Figure 12

Case Investigation 1 – Validation – Title Search – Data.gov

The screenshot shows the Data.gov search results for the query "Mental Health Care in the Last 4 Weeks". The search bar at the top contains the query, and the "Order by" dropdown is set to "Relevance". Below the search bar, the "Organizations" section lists "U.S. Department of Health & Human Services". The "Filter by location" section includes a map of the world and a list of topics, with "Older Adults..." selected. The "Topic Categories" section shows "There are no Topic Categories that match this search". The "Dataset Type" section shows "There are no Dataset Type that match this search". The "Tags" section is empty. The main results area displays three datasets, all marked as "Federal":

- 171 datasets found for "Mental Health Care in the Last 4 Weeks"**
Mental Health Care in the Last 4 Weeks [855 recent views](#)
U.S. Department of Health & Human Services — The U.S. Census Bureau, in collaboration with five federal agencies, launched the Household Pulse Survey to produce data on the social and economic impacts of...
CSV RDF JSON XML
- Telemedicine Use in the Last 4 Weeks** [91 recent views](#)
U.S. Department of Health & Human Services — To rapidly monitor recent changes in the use of telemedicine, the National Center for Health Statistics (NCHS) and the Health Resources and Services Administration's...
CSV RDF JSON XML
- Indicators of Reduced Access to Care Due to the Coronavirus Pandemic During Last 4 Weeks** [16 recent views](#)
U.S. Department of Health & Human Services — The U.S. Census Bureau, in collaboration with five federal agencies, launched the Household Pulse Survey to produce data on the social and economic impacts of...
CSV RDF JSON XML
- NHSC Jobs Center for Primary Care Medical, Dental and Mental Health Providers**
U.S. Department of Health & Human Services — The National Health Service Corps (NHSC)

Note: Centers for Disease Control. (2021). Mental health care in the last 4 weeks [Dataset].

<https://healthdata.gov/dataset/Mental-Health-Care-in-the-Last-4-Weeks/kfr8-6xqg>

Selecting the targeted result “Mental Health Care in the Last 4 Weeks” opened the page where a researcher could download the dataset with different options. In this instance, data validation was successful, and the researcher linked to the actual dataset. This is shown in Figure 13 - Case Investigation 1 – Validation - Title Search - Data.gov displayed below.

Figure 13

Case Investigation 1 – Validation - Title Search - Data.gov

The screenshot shows the Data.gov interface for the dataset 'Mental Health Care in the Last 4 Weeks'. The page header includes the U.S. Department of Health & Human Services logo and a 'Contact Data.gov' button. The dataset title is prominently displayed, followed by the metadata update date: 'Metadata Updated: April 15, 2023'. A detailed description of the survey is provided, explaining its purpose and methodology. Below the description, there is a section for 'Access & Use Information' which states the dataset is public and provides a link to the license. The 'Downloads & Resources' section lists two download options: 'Comma Separated Values File' (with 103 views) and 'RDF File', each with a 'Download' button. A sidebar on the left contains navigation links for 'Topics' (Older Adults Health Data Collection), 'Publisher' (Centers for Disease Control and Prevention), and 'Contact' (National Center for Health Statistics).

Note: Centers for Disease Control. (2021). Mental health care in the last 4 weeks [Dataset].

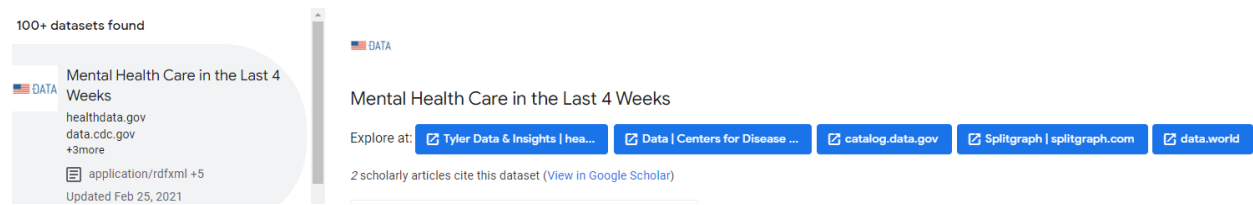
<https://healthdata.gov/dataset/Mental-Health-Care-in-the-Last-4-Weeks/kfr8-6xqg>

A review of the metadata showed it was last updated was April 15, 2023, and the most current search occurred November 7, 2023. Selecting the “Comma Separated Values File” option downloaded the file. The validation for Case Investigation 1 was a pass. As such, this dataset can be identified accurately as the source of the truth to which the Google Dataset Search platform can be compared.

Using the same search criteria for Google Dataset Search returned diverse results because the metadata tags were different. The last update for Google Dataset Search took place on February 25, 2021 (Figure 14).

Figure 14

Case Investigation 1 – Title Search – Validation - Google Dataset Search

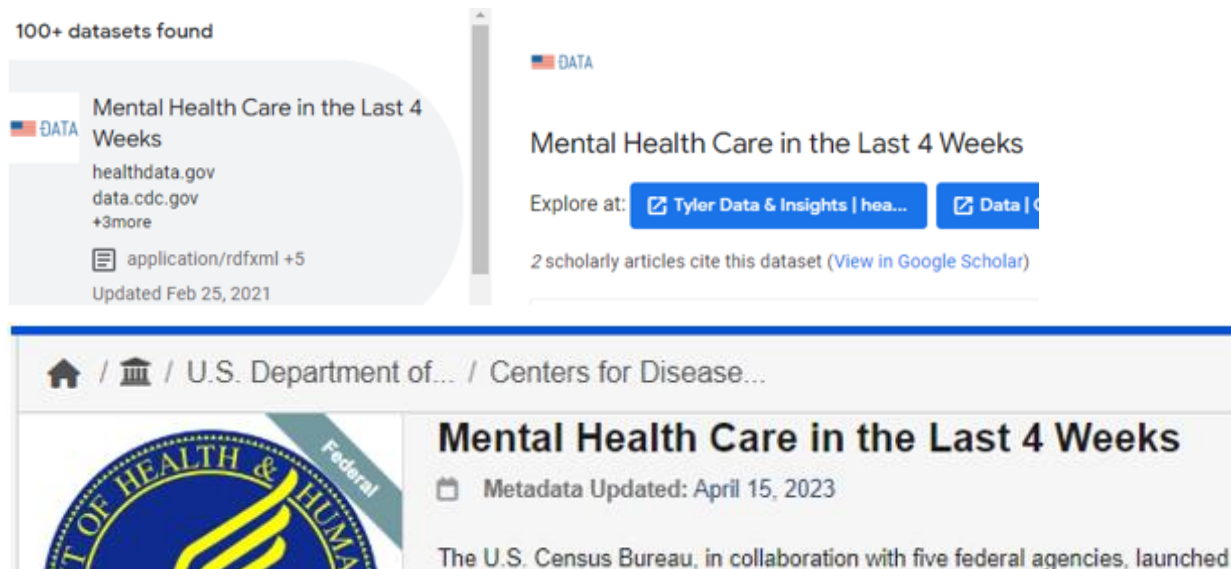


To further verify if this was the same dataset returned from Data.gov, the researcher relied on several tabs Google Dataset Search provided. Five of these tabs, called “Explore Links” were listed. “Explore Links” tabs are hyperlinks to the dataset posted on other sites. One of these is catalog.data.gov – which, when selected, took the researcher to the targeted dataset on Data.gov. The appearance of two differing metadata update dates begs a question: Which of the five links was most likely to match the source of the truth? In this instance, knowledge of the link to catalog.data.gov hinted that it was the valid dataset.

In Case Investigation 1 – Title Search, both platforms featured the valid search target “Mental Health Care in the Last 4 Weeks.” However, questions remain as to the number of links researchers can pursue to find this data, should they use Google Dataset Search. A researcher would have to select each link in Google Dataset Search to validate the dataset. Disparate dates for metadata was an issue noted on this example. The metadata on Google Dataset Search shows it was last updated Feb. 25, 2021. However, in Data.gov, the last update date is shown as April 15, 2023. This discrepancy is noted in Figure 15.

Figure 15

Metadata Discrepancies – Case Investigation 1- Validation



Note: Centers for Disease Control. (2021). Mental health care in the last 4 weeks [Dataset].

<https://healthdata.gov/dataset/Mental-Health-Care-in-the-Last-4-Weeks/kfr8-6xqg>

Overall, the validation was a pass, but there were metadata discrepancies regarding metadata dates updates.

Case Investigation 2 – Keyword Search

The overview of Case Investigation 2 – Keyword Search, is also based on the two platforms. It used specific keywords— “public,” “school,” and “nces”—as search parameters. After a review of results, a random dataset entitled “Public School Characteristics—Current” was selected as the target. Next, the same keywords were input into Google Dataset Search. Ideally, it would return the same target identified in Data.gov. The results for Phase 1 - Evaluating Dataset Findability Case Investigation 2 – Keyword search is displayed in Table 13 below.

Table 13*Case Investigation 2 – Keyword Search Results*

Case 2 Investigation – Keyword Search	Data.gov	Google Dataset Search
Goal:	Find a dataset from the Department of Education using keyword search	Find a dataset from the Department of Education using keyword search
Platform:	Data.gov	Google Dataset Search
Search Query:	“public”, “school”, “nces”	“public”, “school”, “nces”
Parameter:	Keyword	Keyword
Results:	Public School Characteristics – Current	Public School Characteristics – Current
Findable (Yes/No)	Yes	Yes
Number of Results Returned:	13,326	99
Ranking of Dataset:	2	5
Accessible (Yes/No)	Yes – Based on Platform Accessibility	Yes – Based on Platform Accessibility
Interoperable (Yes/No)	Yes – Based on Platform Interoperability	Yes – Based on Platform Interoperability
Reusable (Yes/No)	Yes – Based on Platform Criteria	Yes – Based on Platform Criteria

Case Investigation 2 –Keyword search showed that the platforms using three keywords “public,” “school,” and “nces” returned the targeted dataset – “Public School Characteristics – Current.” Data.gov returned 13,326, but the targeted dataset was ranked number 2. Google

Dataset Search returned only 99 record sets; the ranking of the targeted dataset was five. Such an increase of results is suboptimal, regardless of whether the targeted result set is in the first 10 results. Research has indicated that users are reluctant to scroll past 10 web page results (Wolfram, 2008).

As in Case Investigation 1 – Title Search, the visual aspect of the FAIR compliance criteria becomes evident when examining the Accessible, Interoperable, and Reusable dimensions. By adhering to current platform standards, researchers visually confirm their ability to access the dataset through Data.gov and Google Dataset Search. Additionally, the option to choose from various file formats for downloading the dataset signifies compliance with interoperability standards. Furthermore, the fact that researchers can locate and access the dataset implies reusability.

In Phase 2 – FAIR Compliance Assessment, the actual metadata from each platform was gathered. A worksheet containing numerous detailed questions regarding the dataset outcomes utilized in Case Investigation 2 – Keyword Search is presented in Table 14.

Table 14

Case Investigation 2 – FAIR Compliance Assessment

FAIR Principle	FAIR Compliance Questions	Data.gov DCAT	Google Dataset Search schema.org
Findability	DOI or Globally Unique Identifiers (GUIDs) assigned to the dataset?	Yes	No
Findability	Titles to improve discoverability?	Yes	Yes
Findability	Description to improve discoverability?	Yes	Yes

FAIR Principle	FAIR Compliance Questions	Data.gov DCAT	Google Dataset Search schema.org
Findability	Keyword(s) to improve discoverability?	Yes	No
Accessibility	Registration/Index information in a searchable resource?	Yes	No
Accessibility	Information on how to access the data, including credentials URLs, or APIs?	Yes	Yes
Accessibility	Information on where the data will be stored?	No	No
Accessibility	Information on how long the data will be stored?	No	No
Accessibility	Under what conditions can the data be accessed, including restrictions or licenses listed?	Yes	No
Interoperability	What data formats available listed?	Yes	Yes
Interoperability	What vocabularies or ontologies listed?	Yes	No
Interoperability	What relationships are included listed?	No	No
Reusability	How can the data be used, including restrictions or obligations listed?	Yes	No
Reusability	Where did the data originate listed?	Yes	No
Reusability	What methodology was used to collect the data listed?	No	No

FAIR Principle	FAIR Compliance Questions	Data.gov DCAT	Google Dataset Search schema.org
Reusability	Adherence to community standards and best practices. Is documentation available?	No	No

Of the 16 questions posed by the FAIR compliance framework, Data.gov’s metadata provided affirmative responses to 11. The FAIR Principles areas where Data.gov’s metadata was deficient included Accessibility, Interoperability, and Reusability. The Accessibility guideline criteria for the dataset’s storage and storage duration were not provided. Additionally, Data.gov failed to offer information on the relationships included with the datasets and the methodology used for data collection. Google Dataset Search’s metadata answered affirmatively to six of the 16 questions. Metadata provided by Google Dataset Search included titles, descriptions, and access information through URLs, but did not provide globally unique identifiers, keywords for discoverability, or registration/index information in a searchable resource. Moreover, it did not address accessibility conditions, storage details, vocabulary and ontologies used, data origin, usage restrictions, or documentation adhering to community standards and best practices.

The next step was to evaluate data curation, and this was accomplished by having metadata from the result dataset answer questions. As seen in Table 15 – Case Investigation 2 – Keyword Search – Data Curation Analysis, the metadata has been analyzed.

Table 15

Case Investigation 2 – Keyword Search – Data Curation Analysis

FAIR Principle	Data Curation Questions	Data.gov DCAT	Google Dataset Search schema.org
Findability	Title provides a clear concise summary of the data contents?	Yes	Yes

FAIR Principle	Data Curation Questions	Data.gov DCAT	Google Dataset Search schema.org
Findability	Description provides a clear concise summary of the data contents?	Yes	Yes
Findability	Keywords facilitate search and discovery?	Yes	No
Findability	Creators/Authors provide identification for the person(s) responsible for the data?	Yes	No
Findability	Updated Metadata updated date?	Yes	Yes
Findability	Publication date provides when the data released or published?	Yes	No
Findability	Subject area is the field of study or domain of the data?	No	No
Accessibility	Rights and Licensing info about copyright, usage rights, or other licenses?	Yes	No
Accessibility	Access Restrictions or conditions on accessing data?	Yes	No
Accessibility	Preservation information for data storage, backup procedures, long-term preservation plans?	No	No
Interoperability	File formats the data is stored in?	Yes	Yes
Interoperability	Data Structure description, schema or model?	No	No
Interoperability	Version information?	No	No
Interoperability	Software requirements needed to access the data?	No	No

FAIR Principle	Data Curation Questions	Data.gov DCAT	Google Dataset Search schema.org
Reusability	Source where the data originated from?	Yes	No
Reusability	Collection methods, including how it was collected, including instruments or techniques?	No	No
Reusability	Transformations or processing the data underwent?	Yes	No
Reusability	Citation instructions?	No	No
Reusability	Related publications based or related to dataset?	No	No
Reusability	Use Case example or case studies on how the data has been or can be used?	No	No
Reusability	Publisher?	Yes	No

In the Data Curation Analysis phase, Data.gov provided affirmative answers to 11 of the 21 metadata-based questions. Google Dataset Search answered just four, with particular gaps in all FAIR Principles criteria. It was unable to provide metadata on keywords, creator, publication date, and subject area. As such, it answered only three of seven questions for Findability. In the FAIR Principle – Accessibility, Google Dataset Search it provided no answers to the three questions on licensing, access restrictions, or preservation information on the dataset. Interoperability answers from Goggle Dataset Search provided only the File format; it offered no answers about data structure, versioning, and software requirements. The last section of the data

curation worksheet was based on the Reusability FAIR Principles; Google Dataset Search provided answers to none of those seven questions.

Data.gov answered six of seven metadata questions for Findability and two of three questions regarding Accessibility. Data.gov could not provide preservation information on the dataset. It also could not provide Interoperability answers for data structure, version information, and software requirements. In the Reusability section, three of seven questions were answered. Origination source, transformations performed on the dataset, and publisher answers were provided, but collection methods, citation instructions, related publications, and use case examples were not.

Upon concluding Case Investigation 2, which focused on Keyword Search, the thesis progressed to the validation phase. This step assessed the accuracy and relevance of the identified dataset. The validation phase scrutinized the dataset for quality, consistency, and reliability to authenticate findings.

Case Investigation 2 – Keyword Search – Validation

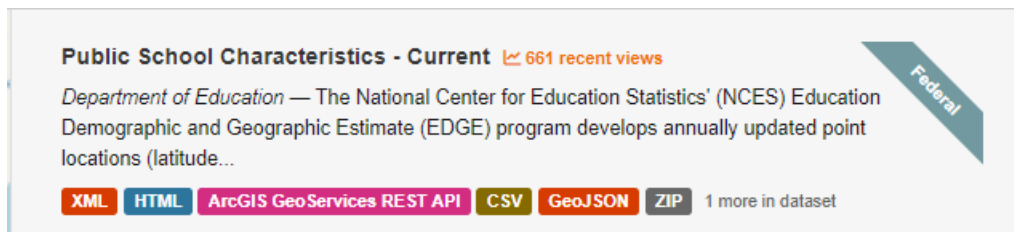
Dataset validation entails a thorough examination of multiple facets of data to confirm its appropriateness and dependability. This process encompasses cross-validation, a method of verification that this thesis uses. In this case study investigation, Data.gov was regarded as the authoritative benchmark, with its datasets serving as the principal reference point.

“Case Investigation 2 – Keyword Search” involved entering the keywords “public,” “school,” and “nces” into a Data.gov search field, yielding a multitude of dataset results, from which one was designated as the focal dataset. This search generated 13,326 dataset outcomes. The chosen dataset, titled “Public School Characteristics – Current,” appeared as the second listing in the search results. The metadata indicated that the most recent update to this dataset

was on November 4, 2023. Navigating to Data.gov and entering “Public School Characteristics – Current” brings the researcher to the following page shown in Figure 16.

Figure 16

Case Investigation 2 – Keyword Search – Validation - Data.gov



Note: National Center for Education (NCES). (2024). Public school characteristics - current [Dataset]. <https://catalog.data.gov/dataset/public-school-characteristics-current-340b1>

Upon selection of “Public School Characteristics – Current,” the user was directed to the landing page (Figure 17) for this dataset, where the researcher could download the file.


Figure 17

Case Investigation 2 – Keyword Search – Validation – Landing - Data.gov

DATA CATALOG

[Home](#) / [Datasets](#) [Organizations](#)

[Home](#) / [Department of Education](#) / [National Center for...](#) [Contact Data.gov](#)



Department of Education

There is no description for this organization

Publisher

National Center for Education Statistics (NCES)

Contact

Andrea Conver

Share on Social Sites

Public School Characteristics - Current

Metadata Updated: November 4, 2023

The National Center for Education Statistics' (NCES) Education Demographic and Geographic Estimate (EDGE) program develops annually updated point locations (latitude and longitude) for public elementary and secondary schools included in the NCES Common Core of Data (CCD). The CCD program annually collects administrative and fiscal data about all public schools, school districts, and state education agencies in the United States. The data are supplied by state education agency officials and include basic directory and contact information for schools and school districts, as well as characteristics about student demographics, number of teachers, school grade span, and various other administrative conditions. CCD school and agency point locations are derived from reported information about the physical location of schools and agency administrative offices. The point locations and administrative attributes in this data layer represent the most current CCD collection. For more information about NCES school point data, see: <https://nces.ed.gov/programs/edge/Geographic/SchoolLocations>. For more information about these CCD attributes, as well as additional attributes not included, see: <https://nces.ed.gov/ccd/files.asp>.Notes:

-1 or M

Indicates that the data are missing.

-2 or N


Indicates that the data are not applicable.

Access & Use Information


[Public](#): This dataset is intended for public access and use.

[License](#): us-pd


Downloads & Resources

Original ISO-19139 metadata


Download

ArcGIS Hub Dataset

Visit page

ArcGIS GeoService

Download

CSV

55 views

Download

Note: National Center for Education (NCES). (2024). Public school characteristics - current [Dataset]. <https://catalog.data.gov/dataset/public-school-characteristics-current-340b1>

Selection of the CSV download produced an error message:

Selecting the download link – a “server error” is displayed. The exact error:

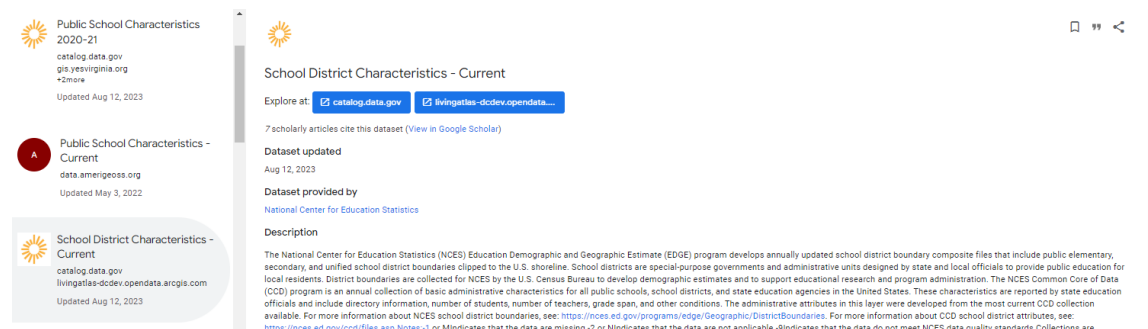
```
{"errors":[{"title":"Server Error","status":500,"message":""}], "meta":{}}
```

Data.gov featured no alternative actions to retrieve this file.

Using the same search parameters, Google Dataset Search returned several records; the targeted dataset was number five. As shown below in Figure 18, the targeted dataset showed that the metadata was last updated August 12, 2023.

Figure 18

Case Investigation 2 – Keyword Search – Validation - Google Dataset Search



Note: National Center for Education (NCES). (2024). Public school characteristics - current [Dataset]. <https://catalog.data.gov/dataset/school-district-characteristics-current-4aa03>

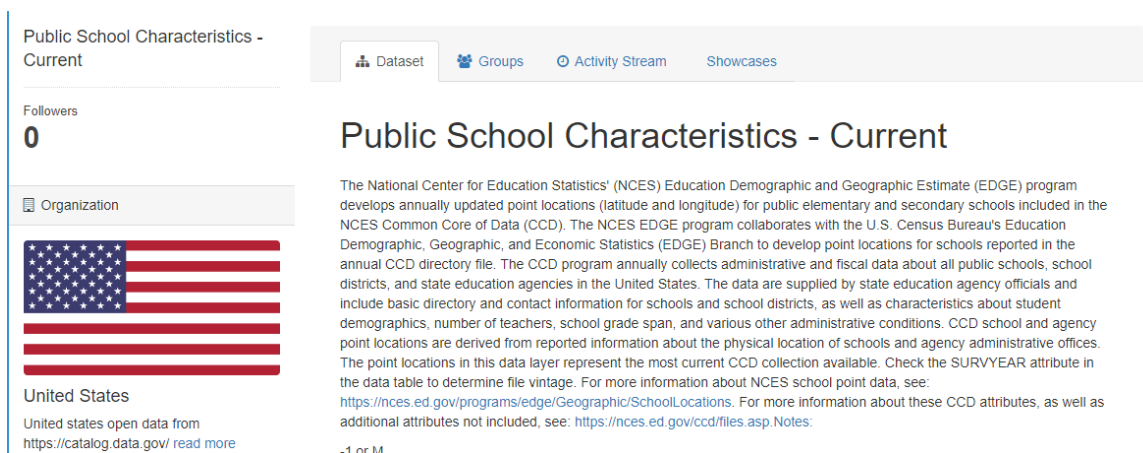
Additional Explorer Links showed catalog.data.gov to be a hyperlink, enabling the researcher to verify it was the correct dataset. Selecting the hyperlink brought the user to the Data.gov platform, and to that specific dataset. Once again, it noted that the last Data.gov update was November 4, 2023. The metadata “create date” in Data.gov was August 12, 2023, indicating that it was, in fact, the correct dataset. Selecting the Download CSV displayed the same error as earlier:

```
{"errors":[{"title":"Server Error","status":500,"message":""}], "meta":{}}
```

Google Dataset Search produced other results showing that the same data was being hosted on other sites. Selecting a different site to verify data from Google Dataset Search allowed the researcher to download a targeted file (Figure 19). However, questions arose about the dataset, including an update date listed as May 3, 2022 (compared to Data.gov, which listed it as November 4, 2023). Additionally, since the original file from Data.gov could not be downloaded, there was no way to compare it to the file selected from the alternate site. The Google Dataset Search site indicated that the data came from Data.gov.

Figure 19

Case Investigation 2 – Keyword Search – Google Dataset Search – Alternative Site



Note: Amerigeo.org. (2024). Public school characteristics - current [Dataset].

<https://data.amerigeoss.org/dataset/public-school-locations-current-282e8>

Case Investigation 2 features validation problems. First, the dataset could not be accessed via download from either “source of the truth.” Data.gov and Google Dataset Search pointed to Data.gov as the link to the dataset. However, selecting it generated the “Server Error.” The second problem was that while Google Dataset Search featured a link to a different site that seemed to host the actual dataset for “Public School Characteristics – Current,” there was no way

to compare the data to Data.gov. This could leave the dataset’s provenance in doubt. Moreover, corruption found in the original source—Data.gov—had been disseminated to the Google Dataset Search platform. Data verification, as an element of sound data curation, is needed to be part of dataset import, export, or aggregation.

Case Investigation 3 – Subject Search

In Case Investigation 3 – Subject Search, a subject was entered as the search parameter in the platforms. A target dataset was selected, and the results returned and ranked based on the targeted result. This investigation used the subject “National Student Loan.” Within this subject, the selected target identified from Data.gov was “National Student Loan Data System.” This dataset was also part of the US Department of Education. The Subject search results are displayed in Table 16.

Table 16

Case Investigation 3 – Subject Search Results

Case 3 Investigation – Subject Search	Data.gov	Google Dataset Search
Goal:	Find the most recent US Government data on the topic “National Student Loan”	Find the most recent US Government data on the topic “National Student Loan”
Platform:	Data.gov	Google Dataset Search
Search Query:	National Student Loan Data	National Student Loan Data
Parameter:	Subject	Subject
Results:	National Student Loan Data	National Student Loan Data
Findable (Yes/No)	Yes	Yes

Case 3 Investigation – Subject Search	Data.gov	Google Dataset Search
Number of Results:	789	100+
Ranking of Dataset:	1	1
Accessible (Yes/No)	Yes – Based on Platform Accessibility	Yes – Based on Platform Accessibility
Interoperable (Yes/No)	Yes – Based on Platform Interoperability	Yes – Based on Platform Interoperability
Reusable (Yes/No)	Yes – Based on Platform Criteria	Yes – Based on Platform Criteria

The results of Case Investigation 3 show that for both platforms, Data.gov and Google Dataset Search, the targeted result “National Student Loan Data System” was returned as a result dataset and was ranked 1. Data.gov returned 789 results while Google Dataset Search returned more than 100. As noted, Findability was rated as “Yes,” as was Accessibility, Interoperability, and Reusability. This occurred because both platforms convey the perception—based on visual cues—that this dataset complies with FAIR compliance criteria.

To provide further analysis, the FAIR Compliance Assessment worksheet was populated with metadata that answered 16 questions relevant to the FAIR Principles. The result of this analysis is shown in Table 17.

Table 17*Case Investigation 3 – Subject Search – FAIR Compliance Assessment*

FAIR Principle	FAIR Compliance Questions	Data.gov DCAT	Google Dataset Search schema.org
Findability	DOI or Globally Unique Identifiers (GUIDs) assigned to the dataset?	Yes	No
Findability	Titles to improve discoverability?	Yes	Yes
Findability	Description to improve discoverability?	Yes	Yes
Findability	Keyword(s) to improve discoverability?	Yes	No
Accessibility	Registration/Index information in a searchable resource?	Yes	No
Accessibility	Information on how to access the data, including credentials URLs, or APIs?	Yes	Yes
Accessibility	Information on where the data will be stored?	No	No
Accessibility	Information on how long the data will be stored?	No	No
Accessibility	Under what conditions can the data be accessed, including restrictions or licenses listed?	Yes	No
Interoperability	What data formats available listed?	Yes	Yes
Interoperability	What vocabularies or ontologies listed?	No	No

FAIR Principle	FAIR Compliance Questions	Data.gov DCAT	Google Dataset Search schema.org
Interoperability	What relationships are included listed?	No	No
Reusability	How can the data be used, including restrictions or obligations listed?	Yes	No
Reusability	Where did the data originate listed?	Yes	No
Reusability	What methodology was used to collect the data listed?	No	No
Reusability	Adherence to community standards and best practices. Is documentation available?	No	No

The FAIR Compliance Assessment analysis for Case Investigation 3 - Subject Search, provided a comprehensive look at FAIR compliance for the dataset “National Student Loan Data System.” For Data.gov, all four questions related to Findability were answered affirmatively. Three of five questions related to Accessibility were also answered. The two unanswered concerned how long the dataset will be stored and where it will be stored. Data.gov answered one of three Interoperability questions. Queries about vocabulary and relationships to other data were not answered. For the Reusability-focused questions, two of four answers were provided. The methodology and community standards and best practices questions were left unanswered.

Google Dataset Search answered four of 16 questions. For Findability, only two of the four questions featured metadata. Only one of five questions on Accessibility was answered; it pertained to the URL location for the dataset. No vocabulary and relationship answers were provided in the Interoperability questions. There were no answers provided in the Reusability

section of the worksheet. Nor were there answers to questions about methodology, dataset origins, or usage restrictions. And no documentation was provided.

Data curation analysis was performed on the dataset “National Student Loan Data System.” The metadata results are illustrated in Table 18.

Table 18

Case Investigation 3 - Subject Search – Data Curation Analysis

FAIR Principle	Data Curation Questions	Data.gov DCAT	Google Dataset Search schema.org
Findability	Title provides a clear concise summary of the data contents?	Yes	Yes
Findability	Description provides a clear concise summary of the data contents?	Yes	Yes
Findability	Keywords facilitate search and discovery?	Yes	No
Findability	Creators/Authors provide identification for the person(s) responsible for the data?	Yes	No
Findability	Updated Metadata updated date?	Yes	Yes
Findability	Publication date provides when the data released or published?	No	No
Findability	Subject area is the field of study or domain of the data?	No	No
Accessibility	Rights and Licensing info about copyright, usage rights, or other licenses?	Yes	No
Accessibility	Access Restrictions or conditions on accessing data?	Yes	No

FAIR Principle	Data Curation Questions	Data.gov DCAT	Google Dataset Search schema.org
Accessibility	Preservation information for data storage, backup procedures, long-term preservation plans?	No	No
Interoperability	File formats the data is stored in?	Yes	Yes
Interoperability	Data Structure description, schema or model?	No	No
Interoperability	Version information?	No	No
Interoperability	Software requirements needed to access the data?	No	No
Reusability	Source where the data originated from?	Yes	No
Reusability	Collection methods, including how it was collected, including instruments or techniques?	No	No
Reusability	Transformations or processing the data underwent?	No	No
Reusability	Citation instructions?	No	No
Reusability	Related publications based or related to dataset?	No	No
Reusability	Use Case example or case studies on how the data has been or can be used?	No	No
Reusability	Publisher?	Yes	No

Each question in the worksheet for data curation is related to a FAIR Principle. There were 21 questions: seven related to Findability; three to Accessibility; four to Interoperability;

and seven to Reusability. The platform Data.gov answered 10 of the 21 questions affirmatively, meaning metadata was available that answered the questions. Data.gov provided five answers regarding Findability, but no publication date and no subject metadata was provided. For Accessibility, two of three questions were posed, with no information about preservation of the dataset provided. However, information on licensing and access restrictions was available. Only one question had an affirmative response in the Interoperability section; it regarded data format. No information on data structure, versioning, or software requirements was found. In the Reusability portion of the worksheet, two of the seven questions answered related to the source origination and publisher. Answers about collection methods, data transformation, citation instructions, related publications, and use case examples were not found.

Focusing on Google Dataset Search, only five of the 21 questions were answered. In Findability, four of the seven questions were answered affirmatively. This included title, description, creator, and updated metadata date. Answers were not available for keywords, publication date, and subject area. Not one of the three questions about Accessibility was answered. There was no metadata for the dataset licensing, access restrictions, or preservation information. Google Dataset Search provided only one answer in Interoperability, which was for file format. No information for data structure, version information, and software requirements was available. Of the seven questions related to Reusability, none were answered. No information was available on the data source, collection methods, transformations, citation instructions, related publications, use case examples, or publisher.

Completion of Case investigation 3 led to the next validation step. Dataset validation is a comprehensive process that examines various data aspects to ensure its suitability and reliability. It includes cross-validation, which this thesis examines as a verification method. The case

investigations considered Data.gov to be the source of the truth and used its data as the primary source.

Case Investigation 3 – Subject Search – Validation

The third case investigation used “Subject” as the search parameter. The target subject was “National Student Loan” from the U.S. Department of Education, which was entered into Data.gov search. Then, the targeted dataset was selected as “National Student Loan Data System.” To ensure dataset validity, the researcher navigated to the page. (See Figure 20 for results.)

Figure 20

Case Investigation 3 – Subject Search – Validation – Data.gov

The screenshot shows the Data.gov dataset page for the "National Student Loan Data System". The page is organized into a sidebar on the left and a main content area on the right. The sidebar includes the ED.gov logo, the Department of Education name, a note about the organization's description, the publisher (Office of Federal Student Aid (FSA)), contact information (Tara Marini), and social media links (Twitter, Facebook). The main content area features the dataset title, a metadata update date (August 12, 2023), a description of the National Student Loan Data System (NSLDS), and sections for "Access & Use Information" (Public access, Creative Commons CCZero license) and "Downloads & Resources" (three downloadable files: 1617FedSchoolCodeList.xlsx, FL_Dashboard_AY_2009_2010_Q1.xls, and FL_Dashboard_AY_2009_2010_Q2.xls, each with a "Download" button).

Home / Department of Education / Office of Federal...

ED.gov Federal

Department of Education
There is no description for this organization

Publisher
Office of Federal Student Aid (FSA)

Contact
Tara Marini

Share on Social Sites
Twitter
Facebook

National Student Loan Data System
Metadata Updated: August 12, 2023

The National Student Loan Data System (NSLDS) is the national database of information about loans and grants awarded to students under Title IV of the Higher Education Act (HEA) of 1965. NSLDS provides a centralized, integrated view of Title IV loans and grants during their complete life cycle, from aid approval through disbursement, repayment, deferment, delinquency, and closure.

Access & Use Information
Public: This dataset is intended for public access and use.
License: Creative Commons CCZero

Downloads & Resources

File Name	Description	Action
1617FedSchoolCodeList.xlsx	Federal School Code List	Download
FL_Dashboard_AY_2009_2010_Q1.xls	2009-2010 Award Year FFEL Q1 Quarterly Activity	Download
FL_Dashboard_AY_2009_2010_Q2.xls	2009-2010 Award Year FFEL Q2 Quarterly Activity	Download

Note: Office of Federal Student Aid (FSA). (2023). National student loan data system [Dataset].

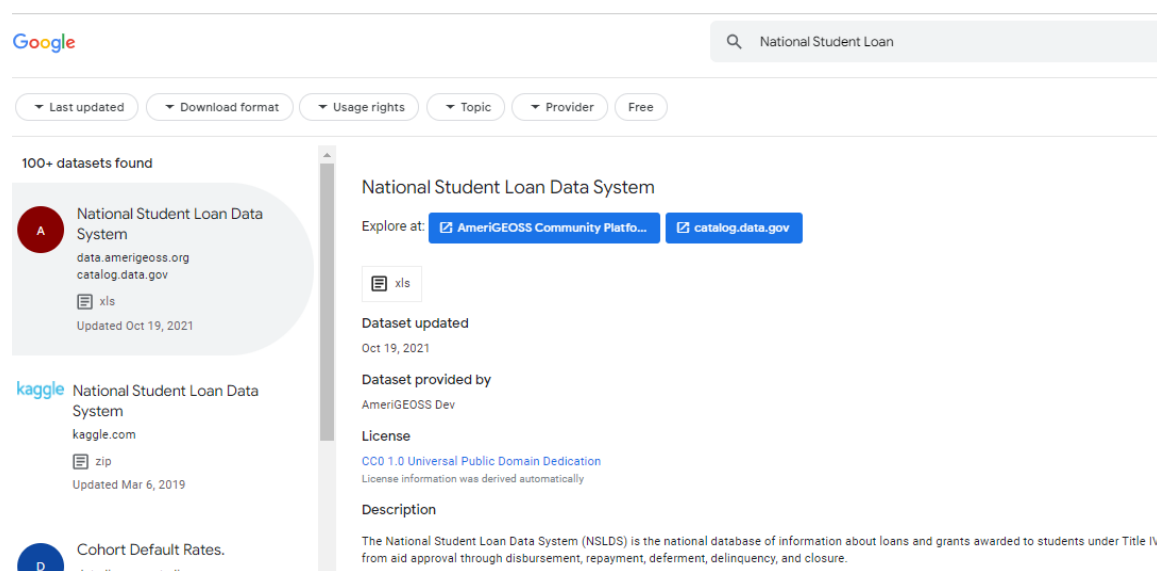
<https://catalog.data.gov/dataset/national-student-load-data-system-722bo>

This page featured 17 possible downloads. The metadata was last updated August 12, 2023. Selection of any of the 17 datasets downloaded an Excel file. In previous testing for this thesis (before August 12, 2023), selection of any of the 17 datasets displayed an error: “This site can’t be reached” or “404 Page not Found.” The researcher was pleased to see Data.gov curators had resolved the errors but left to wonder how they had persisted for so long. The author encountered the error messages in June 2023 and until recently, assumed it was updated and fixed on August 12, 2023, a span of at least two months. As such, data curation was deemed inadequate, and findability compromised. A researcher looking for that information could not obtain it. And, if found on a different site, would it be the same dataset? This led the researcher to question the reliability of data provided by Data.gov.

Using Google Dataset Search, the search parameter “National Student Loan” returned more than 100 results, but the targeted dataset “National Student Loan Data System” was ranked number 1. This is illustrated in Figure 21 below.

Figure 21

Case Investigation 3 – Subject Search – Validation - Google Dataset Search



Note: Office of Federal Student Aid (FSA). (2023). National student loan data system [Dataset].
<https://catalog.data.gov/dataset/national-student-loan-data-system-722b0>

As noted, the metadata update date in Google Dataset Search was October 19, 2021. Using the explorer link, the researcher selected catalog.data.gov and the site Data.gov opened. There was a discrepancy between the metadata update dates. However, as noted earlier, the download links worked correctly. The other explorer link “AmeriGEOSS Community Platform” brought the researcher to a site with identical information regarding the list of 17 datasets. Two sites, Data.gov and AmeriGEOSS, (Figure 22) seemingly featured duplicate datasets. However, the metadata was not identical. As such, unless the researcher compared the downloaded files, there was no reliable way to determine if the dataset was the same. As such, results would not be repeatable if a researcher selected a different dataset. This lack of provenance was potentially problematic.

Figure 22

AmeriGEOSS Site – Case Investigation 3 – Subject Search -Validation

The screenshot displays the AmeriGEOSS website interface. At the top, the AmeriGEOSS logo is on the left, and navigation links for 'Search Datasets', 'DATA HUB', 'DATASETS', 'ORGANIZATIONS', 'GROUPS', and 'ABOUT' are on the right. Below the header, a breadcrumb trail reads 'HOME / ORGANIZATIONS / AMERIGEOSS DEV / NATIONAL STUDENT LOAN DATA SYSTEM'. The main content area is titled 'National Student Loan Data System' and includes a description: 'The National Student Loan Data System (NSLDS) is the national database of information about loans and grants awarded to students under Title IV of the Higher Education Act (HEA) of 1965. NSLDS provides a centralized, integrated view of Title IV loans and grants during their complete life cycle, from aid approval through disbursement, repayment, deferment, delinquency, and closure.' Below this, a 'Data and Resources' section lists five datasets, each with an 'Explore' button: '1617FedSchoolCodeList.xlsx' (Federal School Code List), 'FL_Dashboard_AY_2009_2010_Q1.xls' (2009-2010 Award Year FFEL Q1 Quarterly Activity), 'FL_Dashboard_AY_2009_2010_Q2.xls' (2009-2010 Award Year FFEL Q2 Quarterly Activity), 'FL_Dashboard_AY_2009_2010_Q3.xls' (2009-2010 Award Year FFEL Q3 Quarterly Activity), and 'FL_Dashboard_AY_2009_2010_Q4.xls' (2009-2010 Award Year FFEL Q4 Quarterly Activity). On the left sidebar, there is a 'Followers' section showing 0 followers, an 'Organization' section with the AmeriGEOSS logo and 'Earth Observations for the Americas' tagline, and a 'Social' section for 'AmeriGEOSS Dev' with a 'read more' link.

Note: Amerigeo.org. (2024) National student loan data system [Dataset].
(<https://data.amerigeoss.org/dataset/national-student-loan-data-system>)

The last metadata update was October 20, 2021. When the researcher selected a resource to download, an error appeared: “This site can’t be reached.” It was like the error messages that appeared before August 12, 2023. The unsuccessful download was another example of corruption disseminated from one platform to another. It is also another example of incorrect metadata dates.

Summary

The findings presented in this thesis were designed to explore two related research questions. One was that data curation is integral to improving the FAIR compliance of traditional research datasets (as opposed to being a critical element). However, data format and metadata standardization are more beneficial than traditional data curation (such as organization and categorization). The second research question was the significance of metadata in both curated research datasets and research open datasets. It included the determination as to whether metadata is vital to ensuring the FAIR compliance of data, regardless of whether it is part of a curated dataset or an open dataset.

Open datasets, such as those in Data.gov, are highly curated. They feature standardized metadata and data formats that bolster FAIR compliance compared to Google Dataset Search, which is based on case investigation results from Search types: Title, Keyword, and Subject. While Google Dataset Search returned the required result sets, it necessitated additional attention to ensure the researcher received targeted search results.

The overall FAIR compliance of the three datasets found on the Data.gov platform was illustrated in Table 19. The calculation of the affirmative answer percentage, which reflects the

proportion of completed metadata elements, is achieved using the formula below. An affirmative answer is one in which the question is answered with a provision of complete metadata elements.

The resulting percentage offers an indication of FAIR compliance.

$$\text{Percentage of Complete Metadata Elements} = (\text{Number of Complete Metadata Elements} / \text{Total Number of Questions}) \times 100$$

Table 19

FAIR Compliance Assessment – Data.gov

FAIR Principle	FAIR Compliance Questions – Data.gov	Case Investigation 1- Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Findability	DOI or Globally Unique Identifiers (GUIDs) assigned to the dataset?	Yes	Yes	Yes
Findability	Titles to improve discoverability?	Yes	Yes	Yes
Findability	Description to improve discoverability?	Yes	Yes	Yes
Findability	Keyword(s) to improve discoverability?	Yes	Yes	Yes
Accessibility	Registration/Index information in a searchable resource?	Yes	Yes	Yes
Accessibility	Information on how to access the data, including credentials URLs, or APIs?	Yes	Yes	Yes

FAIR Principle	FAIR Compliance Questions – Data.gov	Case Investigation 1- Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Accessibility	Information on where the data will be stored?	No	No	No
Accessibility	Information on how long the data will be stored?	No	No	No
Accessibility	Under what conditions can the data be accessed, including restrictions or licenses listed?	Yes	Yes	Yes
Interoperability	What data formats available listed?	Yes	Yes	Yes
Interoperability	What vocabularies or ontologies listed?	Yes	Yes	No
Interoperability	What relationships are included listed?	Yes	No	No
Reusability	How can the data be used, including restrictions or obligations listed?	Yes	Yes	Yes
Reusability	Where did the data originate listed?	Yes	Yes	Yes
Reusability	What methodology was used to collect the data listed?	Yes	No	No

FAIR Principle	FAIR Compliance Questions – Data.gov	Case Investigation 1- Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Reusability	Adherence to community standards and best practices. Is documentation available?	No	No	No

Although datasets from Data.gov do not achieve the 100% target for FAIR compliance, they approach it. For instance, the dataset examined in Case Investigation 1, titled “Mental Health Care in the Last 4 Weeks,” achieved a FAIR compliance score of 81.25%. Meanwhile, the dataset involved in Case Investigation 2, found with the keyword search “Public School Characteristics – Current,” reached a FAIR compliance score of 68.75%. Lastly, the dataset for Case Investigation 3, identified through a subject search for the “National Student Loan Data System,” attained a FAIR compliance score of 62.5%.

Alternately, Google Dataset Search attained a FAIR compliance score of 25% for all three datasets analyzed, as noted in Table 20 – FAIR Compliance Assessment – Google Dataset Search.

Table 20

FAIR Compliance Assessment – Google Dataset Search

FAIR Principle	FAIR Compliance Questions – Google Dataset Search	Case Investigation 1- Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Findability	DOI or Globally Unique Identifiers (GUIDs) assigned to the dataset?	No	No	No

FAIR Principle	FAIR Compliance Questions – Google Dataset Search	Case Investigation 1- Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Findability	Titles to improve discoverability?	Yes	Yes	Yes
Findability	Description to improve discoverability?	Yes	Yes	Yes
Findability	Keyword(s) to improve discoverability?	No	No	No
Accessibility	Registration/Index information in a searchable resource?	No	No	No
Accessibility	Information on how to access the data, including credentials URLs, or APIs?	Yes	Yes	Yes
Accessibility	Information on where the data will be stored?	No	No	No
Accessibility	Information on how long the data will be stored?	No	No	No
Accessibility	Under what conditions can the data be accessed, including restrictions or licenses listed?	No	No	No
Interoperability	What data formats available listed?	Yes	Yes	Yes

FAIR Principle	FAIR Compliance Questions – Google Dataset Search	Case Investigation 1- Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Interoperability	What vocabularies or ontologies listed?	No	No	No
Interoperability	What relationships are included listed?	No	No	No
Reusability	How can the data be used, including restrictions or obligations listed?	No	No	No
Reusability	Where did the data originate listed?	No	No	No
Reusability	What methodology was used to collect the data listed?	No	No	No
Reusability	Adherence to community standards and best practices. Is documentation available?	No	No	No

Data.gov demonstrated better FAIR compliance scores in comparison to Google Dataset Search, effectively addressing RQ2: The role of metadata in enhancing the FAIR compliance of curated research datasets versus open datasets. The increased volume of metadata presents in Data.gov correlates with heightened adherence to the FAIR Principles – Findability, Accessibility, Interoperability, and Reusability. This thesis aimed to validate the assumption that metadata elevates the FAIR compliance of both conventional and open research datasets. The expectation was that metadata’s detailed and organized format would promote FAIR compliance

of datasets within curated traditional research environments. For open datasets, the research intended to explore the ways in which uniform metadata protocols could improve access and interoperability, thereby emphasizing their importance in contrast to traditional datasets.

RQ1: What role does data curation play in the FAIR compliance of traditional research datasets and how does this role differ in the context of open datasets? This was answered by evaluating the data curation metadata extracted during Case Investigations 1, 2, and 3. The datasets were queried against specific data curation questions associated to the FAIR Principles and components of data curation. Starting with Data.gov, Table 21 provides an overview of the results of the questions provided for data curation.

Table 21

Data Curation – Data.gov

FAIR Principle	Data Curation Questions -Data.gov	Case Investigation 1 – Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Findability	Title provides a clear concise summary of the data contents?	Yes	Yes	Yes
Findability	Description provides a clear concise summary of the data contents?	Yes	Yes	Yes
Findability	Keywords facilitate search and discovery?	Yes	Yes	Yes
Findability	Creators/Authors provide identification for the person(s)	Yes	Yes	Yes

FAIR Principle	Data Curation Questions -Data.gov	Case Investigation 1 – Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
	responsible for the data?			
Findability	Updated Metadata updated date?	Yes	Yes	Yes
Findability	Publication date provides when the data released or published?	Yes	Yes	No
Findability	Subject area is the field of study or domain of the data?	Yes	No	No
Accessibility	Rights and Licensing info about copyright, usage rights, or other licenses?	Yes	Yes	Yes
Accessibility	Access Restrictions or conditions on accessing data?	Yes	Yes	Yes
Accessibility	Preservation information for data storage, backup procedures, long-term preservation plans?	No	No	No
Interoperability	File formats the data is stored in?	Yes	Yes	Yes

FAIR Principle	Data Curation Questions -Data.gov	Case Investigation 1 – Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Interoperability	Data Structure description, schema or model?	No	No	No
Interoperability	Version information?	No	No	No
Interoperability	Software requirements needed to access the data?	No	No	No
Reusability	Source where the data originated from?	Yes	Yes	Yes
Reusability	Collection methods, including how it was collected, including instruments or techniques	Yes	No	No
Reusability	Transformations or processing the data underwent?	Yes	Yes	No
Reusability	Citation instructions?	No	No	No
Reusability	Related publications based or related to dataset?	No	No	No
Reusability	Use Case example or case studies on how the data has been or can be used?	No	No	No

FAIR Principle	Data Curation Questions -Data.gov	Case Investigation 1 – Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Reusability	Publisher?	Yes	Yes	Yes

The higher the percentage of affirmations provides a better chance of effective data curation processes. In Case Investigation 1 – Title Search, for Data.gov platform, the data curation percentage was 66.7%—the highest of the three Case Investigations. Case Investigation – Keyword Search scored 57.1% and Case Investigation 3 – Subject Search was lowest at 47.6%.

Conversely, Google Dataset Search had much lower scores for data curation. Table 22 – Data Curation – Google Dataset Search shows that metadata elements needed for data curation processes are missing.

Table 22

Data Curation – Google Dataset Search

FAIR Principle	Data Curation Questions -Google Dataset Search	Case Investigation 1 – Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Findability	Title provides a clear concise summary of the data contents?	Yes	Yes	Yes
Findability	Description provides a clear concise summary of the data contents?	Yes	Yes	Yes
Findability	Keywords facilitate search and discovery?	No	No	No

FAIR Principle	Data Curation Questions -Google Dataset Search	Case Investigation 1 – Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Findability	Creators/Authors provide identification for the person(s) responsible for the data?	Yes	No	Yes
Findability	Updated Metadata updated date?	Yes	Yes	Yes
Findability	Publication date provides when the data released or published?	No	No	No
Findability	Subject area is the field of study or domain of the data?	No	No	No
Accessibility	Rights and Licensing info about copyright, usage rights, or other licenses?	No	No	No
Accessibility	Access Restrictions or conditions on accessing data?	No	No	No
Accessibility	Preservation information for data storage, backup procedures, long- term preservation plans?	No	No	No

FAIR Principle	Data Curation Questions -Google Dataset Search	Case Investigation 1 – Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
Interoperability	File formats the data is stored in?	Yes	Yes	Yes
Interoperability	Data Structure description, schema or model?	No	No	No
Interoperability	Version information?	No	No	No
Interoperability	Software requirements needed to access the data?	No	No	No
Reusability	Source where the data originated from?	No	No	No
Reusability	Collection methods, including how it was collected, including instruments or techniques	Yes	No	No
Reusability	Transformations or processing the data underwent?	Yes	No	No
Reusability	Citation instructions?	No	No	No
Reusability	Related	No	No	No

FAIR Principle	Data Curation Questions -Google Dataset Search	Case Investigation 1 – Title Search	Case Investigation 2 – Keyword Search	Case Investigation 3 – Subject Search
	publications based or related to dataset?			
Reusability	Use Case example or case studies on how the data has been or can be used?	No	No	No
Reusability	Publisher?	Yes	No	No

In Case Investigation 1 – Title Search dataset named “Mental Health Care in the Last 4 Weeks,” the percentage of collected data curation elements was 38.1%. This was the highest of the scores for the three Case Investigations. Case Investigation 3 – Subject Search dataset titled “National Student Loan Data System” scored 23.8%. The lowest score of 19% was assigned to Case Investigation 2 – Keyword Search based on the dataset “Public School Characteristics – Current.”

Does this data answer RQ1: What role does data curation play in the FAIR compliance of traditional research datasets? Evaluating the percentage of completeness as it pertains to each FAIR Principle for each Case Investigation is a better way to analyze the answer to this question. Including both platforms, Data.gov and Google Dataset Search.

In Data.gov, the Findability FAIR Principles for Case Investigation 1 – Title Search was 100%, meaning every Findability element was present. For Case Investigation 2 – Keyword Search the score was 85.7%. The final score for Case Investigation 3 – Subject Search was 71.4%. Overall, the impact of data curation for Findability ranked high. The Accessibility FAIR

Principle scored 66.7% for Case Investigation 1, 85.7% for Case Investigation 2, and 71.4% for Case Investigation 3. Interoperability scores for all three Case Investigations were 25%.

Reusability FAIR Principle scored 57.1% for Case Investigation 1; the score for Case Investigation 2 was 42.9%, and for Case Investigation 3, the score was 28.6%. Table 23 – Data.gov – Data Curation Breakdown recaps these numbers below.

Table 23

Data.gov – Data Curation Breakdown

Data.gov	Case Investigation 1 Title Search	Case Investigation 2 Keyword Search	Case Investigation 3 Subject Search
Overall	66.7%	57.1%	47.6%
Findability	100.0%	85.7%	71.4%
Accessibility	66.7%	66.7%	66.7%
Interoperability	25.0%	25.0%	25.0%
Reusability	57.1%	42.9%	28.6%

Conversely, Google Dataset Search provided much lower scores. For Case Investigation 1 – Title Search, the overall score for data curation was 38.1%, while the score for Case Investigation 2 – Keyword Search was 19%. The overall score for Case Investigation 3 – Subject Search was 23.8%. When scores are analyzed according to the individual FAIR Principles, the results reveal a nuanced narrative. This information is shown in Table 24 – Google Dataset Search – Data Curation Breakdown.

Table 24*Google Dataset Search – Data Curation Breakdown*

Google Dataset Search	Case Investigation 1 Title Search	Case Investigation 2 Keyword Search	Case Investigation 3 Subject Search
Overall	38.1%	19.0%	23.8%
Findability	57.1%	42.9%	57.1%
Accessibility	0.0%	0.0%	0.0%
Interoperability	25.0%	25.0%	25.0%
Reusability	42.9%	0.0%	0.0%

In the Findability FAIR Principles Google Dataset Search reported scores of 57.1% for both Case Investigations 1 and 2. Case Investigation 3 reported a 42.9% score. Accessibility scores came in at 0.0%. This was based on the lack of metadata for data curation questions pertaining to the principle. Interoperability across all three Case investigations scored 25%. Finally, the Reusability FAIR Principles report 42.9% for Case Investigation 1 and 0.0% for both Case Investigations 2 and 3.

Based on the overall analysis of data curation-based questions, the answer to RQ1 is that data curation plays little to no role in the FAIR compliance of traditional research datasets or open research datasets. The metadata associated with data curation is based more on the requirements of the platform or repository. For example, Data.gov has standard metadata schemas, but data curation-based metadata is not part of its mandatory standards. Google Dataset Search requires only bare-minimum metadata standards.

The findings from the validation process remain an outstanding discussion topic. Specifically, in Case Investigation 1, which focused on Title Search, a discrepancy was noted between the metadata update dates on Data.gov and Google Dataset Search. Data.gov listed the update as occurring on April 15, 2023, whereas Google Dataset Search noted February 25, 2021. That variance raises questions about the dataset's provenance and which source is more reliable. Despite the conflicting information, since updates were registered on both platforms, the dataset would meet the criteria for FAIR compliance and data curation standards.

The outcomes of the validation process in Case Investigation 2 – Keyword Search are also worth consideration. On Data.gov, the researcher was able to locate the dataset named “Public School Characteristics – Current.” However, when attempting to download the data in CSV format, an error message indicated a server issue, blocking access to the file. There were no alternative means to obtain the dataset on Data.gov. In contrast, the same dataset appeared on Google Dataset Search, which redirected to Data.gov, leading to the same server error. Additional links on Google Dataset Search directed to different sources for the dataset, presenting a mismatch in metadata update dates – the date on Google Dataset Search for this alternate site being May 3, 2022, as opposed to November 4, 2023, on Data.gov. This variation caused uncertainty about the dataset's consistency. Despite these issues, the dataset would still satisfy the FAIR compliance and data curation criteria given its identifiable URL location and documented metadata updates.

Another issue discovered during the validation phase in Case Investigation 3 – Subject Search pertained to corrupt URLs for downloads. On Data.gov, the landing page for the dataset “National Student Loan Data System,” 17 individual datasets were listed for downloads as both xls and xlsx file formats. Selecting any of the 17 datasets led to an error message: “This site can't

be reached” or “404 Page not Found.” Subsequent testing—after August 12, 2023—showed that the issue was resolved. However, for about two months before this date—in June 2023—this data was not available for download. There are other links to this dataset on Google Dataset Search, including one that redirects to Data.gov. However, URL locations for the data downloads point to a URL that looks like Data.gov but is incorrect. Instead, researchers receive a “site can’t be reached” error. Despite these issues, the dataset would still satisfy the FAIR compliance and data curation criteria because its Interoperability file format metadata is correct, and the metadata features a URL download location.

The conclusions drawn in this thesis corroborate the argument that metadata significantly impacts the FAIR compliance of datasets within both traditional and open research contexts. Additionally, the research demonstrates that through data curation, datasets can be thoroughly documented, preserved, and prepared for use and subsequent reuse. However, it’s important to note that data curation practices are influenced by the standards of the platforms that host the datasets, rather than being intrinsic to the datasets themselves.

CHAPTER 5

DISCUSSION, RECOMMENDATIONS, and CONCLUSION

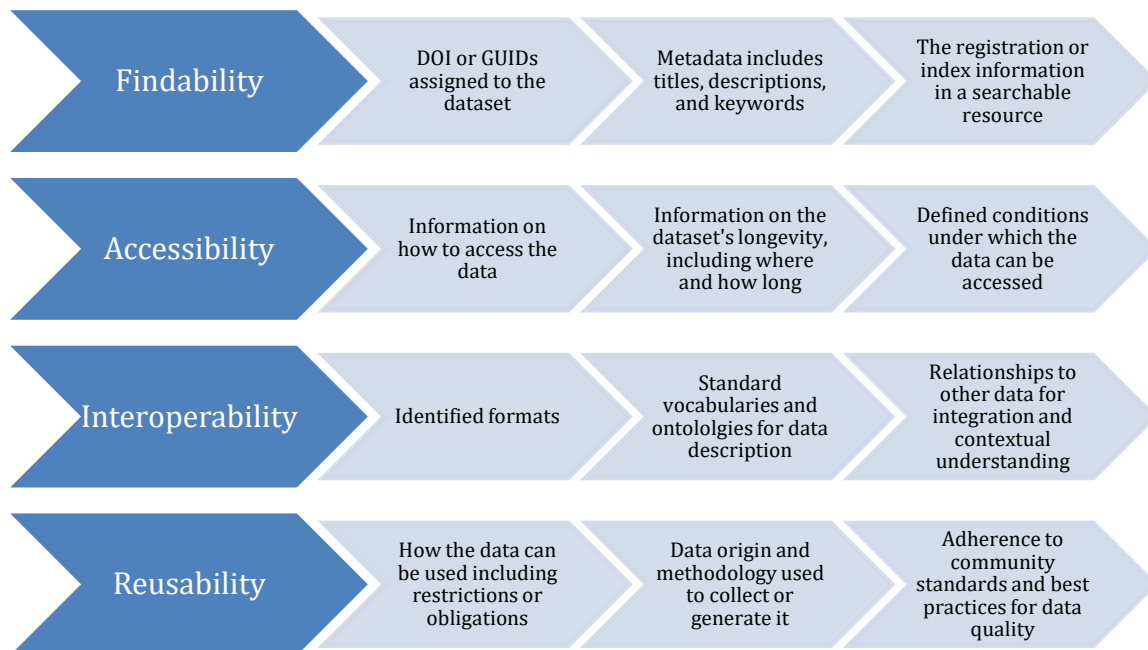
Discussion of the Findings

In this discussion, we present the key conclusions and findings from this thesis that examined data curation's role in the FAIR compliance of traditional research datasets—and how this role differs in research open datasets. This thesis also explored the significance of metadata in curated research datasets and open datasets. The findings discussed pertain to both traditional research datasets and open research datasets.

As part of the findings, the concept of FAIR compliance was introduced and used as a guideline for measuring how “FAIR” was a selected dataset. The criteria for FAIR compliance were described using the following guidelines (see figure 23 below).

Figure 23

FAIR Compliance Criteria



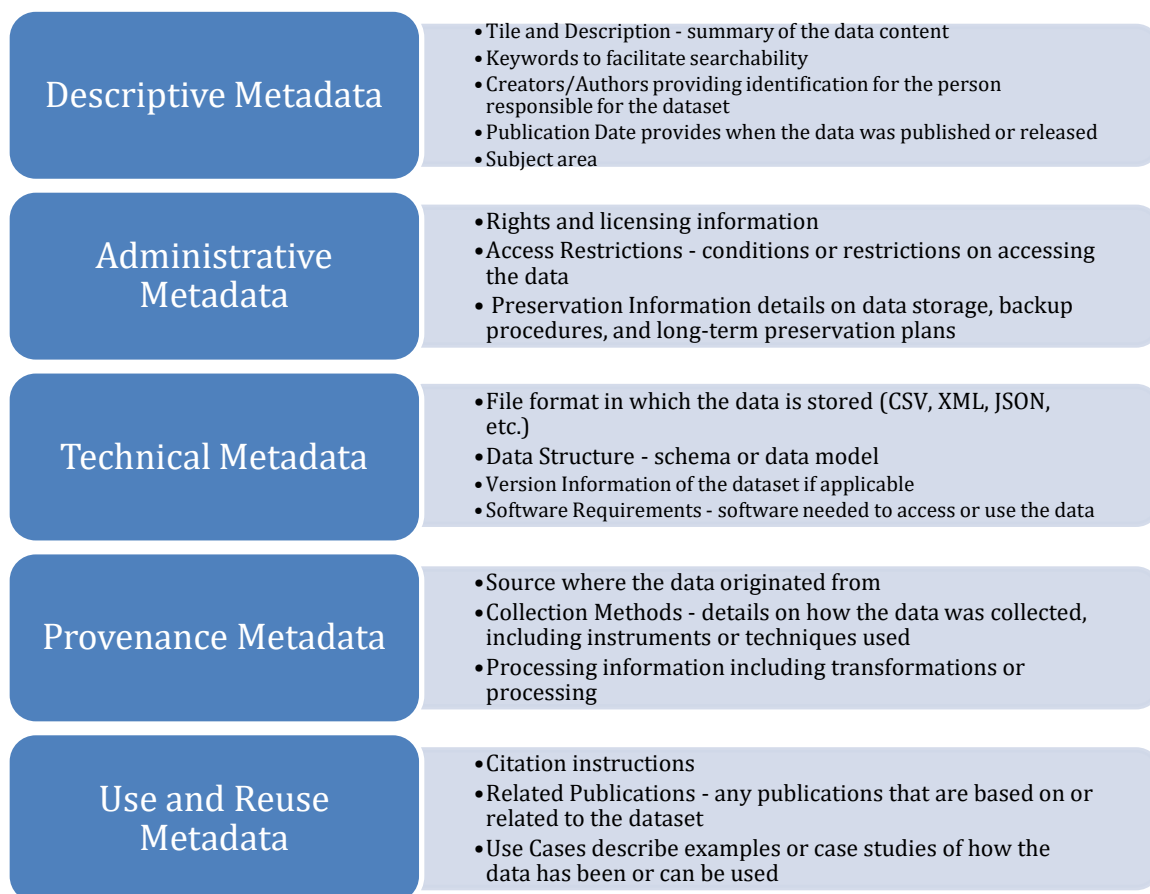
Note. Adapted from "The FAIR guiding principles for scientific data management and stewardship" by Wilkinson et al., 2016, Scientific Data, 3(1), Article

1. (<https://doi.org/10.1038/sdata.2016.18>). In the public domain.

The presented findings demonstrated a need for better data curation methods as it pertains to the platform the data is hosted by or linked to. Data curation guidelines are presented in Figure 24 below.

Figure 24

Data Curation Guidelines



Note. Adapted from: "Understanding metadata" by J. Riley, 2017. p. 7

In both Figure 23 and Figure 24, the criteria for most FAIR compliance and data curation guidelines are contained in the metadata associated with each dataset. As seen in Table 24 (below), there is an overlap between the FAIR Principles guidelines and data curation metadata criteria. Table 24 does not show every possible relationship between the FAIR Principles and data curation metadata, but it does illustrate the direct relationship between the two. The description field in Table 25 explains how elements from both the FAIR Principles and data curation metadata are connected.

Table 25

FAIR Principles Overlap with Data Curation Metadata.

FAIR Principles	Data Curation Metadata	Description
Findability Guidelines	Descriptive Metadata	Both focus on the use of titles, descriptions, and keywords to make data easily discoverable.
Accessibility Guidelines	Administrative Metadata	Information on access conditions, restrictions, and licensing details for the dataset.
Interoperability Guidelines	Technical Metadata	Standards on data formats and structures that facilitate the use and integration of the dataset with other data.
Reusability Guidelines	Use and Reuse Metadata Provenance Metadata	Guidelines and metadata on how the data can be reused, including details on the data's origin, collection methodology, and instructions for citations.

From a technical perspective, if the dataset contains the metadata, that dataset could be considered FAIR and/or the data curation processes successfully completed. The findings in this thesis counteract this assumption. For example, Data.gov mandates that certain keywords, like “Covid-19,” be included in datasets, even if only marginally relevant to a topic. This forced inclusion of keywords can generate a multitude of datasets that do not pertain to Covid-19, confounding researchers’ efforts to identify datasets tightly related to their needs. This specific mandate reduces search precision and can waste researchers’ time as they sift through datasets that are only tangentially related to their actual interests. This practice can also inflate the perceived number of resources related to a topic, resulting in misleading statistics about the availability of relevant data.

Furthermore, metadata elements such as “publisher” and “creation date” are influenced by a user’s perspective, which can introduce subjectivity and bias. A dataset may be published by an organization or individual who interprets its content differently from a researcher. Similarly, the determination of the creation date could vary depending on the user’s viewpoint, such as a dataset lifecycle point in time, like when it was initially gathered, when it was processed, or when it was made available for use. Different treatment of this singular data point could lead to divergent search results and hinder accuracy. Such discrepancies could impact search results and data reliability, as researchers may be presented with datasets that do not align temporally with their research period. Additionally, researchers could miss out on relevant datasets due to inconsistent date reporting. These examples illustrate that while metadata is a fundamental component of the FAIR Principles, the quality, accuracy, and consistency of metadata are equally important.

To address these challenges, it's clear that additional components need to be implemented, but not necessarily at the individual dataset level. The solution lies in enhancing data curation methods through the development of comprehensive policies. I recommend that these policies serve as a deliberate system of principles to guide and achieve rational outcomes, acting as statements of intent implemented as procedures or protocols. By establishing a framework of best practices, such guidelines would be designed to influence or determine decisions and actions across a wide range of issues or domains. Instituting such policies would mitigate the subjective variances in metadata creation and provide a standard that ensures the integrity of and utility of datasets, aligning with the core tenants of the FAIR Principles.

ISO Policy

A policy needs to be based upon a framework and the International Organization for Standardization (ISO) provides that support for a broad range of fields. ISO creates standards to ensure quality, safety, and efficiency. ISO standards are recognized globally, making them more advantageous than other options that may only have regional acceptance or recognition. Global recognition helps ensure that products and services meet international benchmarks, improving creditability and marketability. Numerous policies, across various domains and industries, rely on ISO standards to ensure quality, safety, and efficiency. Table 26 shows some examples of ISO-based policies.

Table 26

ISO Based Policy Examples

Policy	Description
Environmental Management Polices:	Utilize ISO 14001 to minimize environmental impact.

Policy	Description
Quality Management Policies:	Follows ISO 9001 to enhance product and service quality.
Energy Management Policies:	Incorporates ISO 50001 to improve energy efficiency.
Information Security Management Policies:	Adheres to ISO/IEC 27001 to secure information assets.
Health and Safety Management Policies:	Implementing ISO 45001 to ensure workplace safety.
Data Management Policies:	Applying 25012 standards to various aspects of data management, such as classification, storage, and processing to maintain data integrity, accuracy, and relevance

Table 26 as noted above illustrates examples of ISO-based policies, highlighting the extensive reach of the ISO, which has published more than 25,000 standards (*ISO - Standards*, n.d.). Consequently, it is logical to assume that numerous policies across various sectors are derived from these standards, affecting areas such as healthcare, environmental management, technology, manufacturing, and public administration. One example is ISO 25012, a standard utilized in Data Management Policies and software application development. It is globally recognized as a data quality model (Simonetta et al., 2021). According to Gualo et al., (2021) data quality is defined by criteria such as accuracy, completeness, reliability, and relevance, ensuring that data is:

- Accurate: Free of errors and accurately reflects reality or a source.
- Complete: Contains all required data without omissions.

- **Consistent:** Maintains uniformity across different datasets and corresponds with previous records.
- **Timely:** Accessible as needed and represents the most up-to-date conditions.
- **Reliable:** Trustworthy for use in decision-making.
- **Relevant:** Suitable for the contexts in which it is employed.

To develop a data curation and FAIR compliance policy based on the ISO/IEC 25012 standards, it is essential to understand the data quality characteristics the standard defines. These characteristics fall into two categories: Inherent and system-dependent data quality. Inherent data quality assesses how naturally data meets its intended purpose, focusing on the core attributes that fulfill user needs. System-dependent data quality evaluates how the data's quality is supported and maintained by underlying systems and technologies, ensuring data remains accurate, complete, and dependable through specific tools and processes.

ISO 25012 Inherent Characteristics and Alignments

The inherent characteristics of ISO/IEC 25012 are described in Table 27 below. These characteristics, as detailed in the standard, provide the criteria to evaluate and ensure data quality in different domains.

Table 27

Inherent Characteristics of ISO/IEC 25012

Inherent Characteristic	Description
Accuracy	The extent to which data correctly reflects the actual attributes or scenario it is intended to depict.
Completeness	The degree to which all required data is known.

Inherent Characteristic	Description
Consistency	The extent to which the data is consistent, within the dataset and across different datasets.
Credibility	The degree to which the data is true and believable.
Currentness	The degree to which the data is up to date.
Accessibility	The degree to which data can be accessed in a specific context of use.
Compliance	The degree to which the data complies with stated rules and regulations.
Confidentiality	The degree to which the data ensures privacy or confidentiality.

Note: Adapted from "Data quality certification using ISO/IEC 25012: Industrial experiences" by F. Gualo et al., 2021, Journal of Systems and Software (<https://doi.org/10.48550/arXiv.2102.11527>)

The FAIR Principles align with several inherent data quality characteristics described in ISO/IEC 25012 (and as defined in Table 27). Incorporating these characteristics into a policy will ensure datasets meet quality standards. It also guarantees they are managed in ways that maximize their value for a wide range of uses and users over time. Alignment of the FAIR Principles and inherent characteristics of ISO/IEC 25012 is explained in Table 28.

Table 28

ISO/IEC 25012 and FAIR Principles Alignment

ISO/IEC 25012 Characteristic	FAIR Principles Guideline	Alignment
Accessibility	Findability	Emphasizes the importance of data being locatable and findable.

ISO/IEC 25012 Characteristic	FAIR Principles Guideline	Alignment
Accessibility	Accessibility	Ensuring data can be accessed by both humans and machines once its location is known.
Consistency	Interoperability	The data's ability to integrate with other data.
Credibility	Interoperability	Interoperable data is often more credible.
Currentness	Reusability	Ensures data remains applicable for various purposes.
Accuracy	Reusability	Ensures data remains accurate over time.

These inherent characteristics also align with the data curation guidelines defined in this thesis, as shown in Table 29 below. These alignments ensure that data managed under ISO/IEC 25012 standards are well-prepared for effective use and reuse, and supported by comprehensive metadata that facilitates searchability, provenance tracking, and long-term preservation.

Table 29

ISO/IEC 25012 and Data Curation Guidelines Alignment

ISO/IEC 25012 Characteristic	Data Curation Guideline	Alignment
Accuracy Completeness	Descriptive and Provenance metadata	Foundational for ensuring data accurately represents its intended subject and that all necessary information is captured.
Consistency Credibility	Administrative and Technical metadata	Maintains uniformity across datasets and verifying the reliability of data sources.

ISO/IEC 25012 Characteristic	Data Curation Guideline	Alignment
Currentness Accessibility	Administrative metadata	Crucial for ensuring Preservation Information and Access Restrictions in Administrative metadata is up-to-date and datasets are readily available for use.
Compliance Confidentiality	Administrative metadata	Essential for crafting Rights and Licensing, and Access Restrictions in Administrative metadata. Adherence to legal standards and safeguarding sensitive information.

A policy designed for FAIR compliance and data curation can effectively leverage the inherent characteristics of the ISO/IEC 25012 standard to ensure robust data management via metadata. These characteristics intertwine with both FAIR Principles and data curation guidelines, enhancing the utility and integrity of data by utilizing metadata elements—specifically schema.org—which is a universally acknowledged standard:

Accessibility: Use “Access Restrictions” metadata to detail how and under what conditions data can be accessed, ensuring it remains easily retrievable. Schema.org lacks a dedicated element that can detail access to the data. One field that could contain this information is ‘accessibilitySummary.’ The property provides a summary of how accessible the content is.

Consistency: “Data Structure” metadata describes the dataset’s schema or model, helping maintain uniformity across datasets. In schema.org metadata, there is no specific element named “Data Structure,” but several properties that can be used to hold this data. These fields include “sameAs,” “identifier,” “about,” and “url.” These fields can demonstrate the consistency of content or data.

Credibility: “Source” metadata, indicating the data origin, establishes its credibility by confirming validity of data sources. While schema.org does not have a property named “Source,” several alternative properties can establish and convey content credibility. These properties are “author,” “publisher,” “review,” “aggregateRating,” “award,” “citation,” and “isBasedOn.” Any of these could communicate the reliability and authoritative nature of the content to users and search engines.

Currentness: “Publication Date” or “Version Information” metadata tracks when data was last updated or published, ensuring its relevance. Schema.org features the following properties that can indicate “Currentness”: “dateModified,” “datePublished,” “validThrough,” and “validFrom.”

Accuracy: “Validation” metadata, which documents methods to check data accuracy, ensures that data correctly represents real-world conditions. Schema.org does not directly offer a metadata element for “Validation.” However, the property “description” could include details about validation outcomes. Another schema.org element, “softwareVersion,” could indicate it has passed validation.

Completeness: “Completeness” metadata, which specifies whether all required fields are populated, helps assess dataset comprehensiveness. Schema.org lacks an element specifically for “completeness” of data or content. Other general attributes such as “description” or “additionalType” can be used to include details about data completeness. Additionally, the property “about” could be leveraged to specify what the data covers.

Compliance: “Rights and Licensing” metadata provides information about compliance with legal constraints and usage rights. Schema.org does not directly provide a specific metadata

element name “compliance.” However, the ‘description’ metadata element can include compliance details, or the “about” property can specify what the compliance is related to.

Confidentiality: “Confidentiality” metadata outlines measures established to protect sensitive information within the dataset. Schema.org metadata, which Google Dataset Search uses, does not include an element for “Confidentiality,” but the “description” or “name” elements can include this information.

In considering the inherent data quality characteristics defined by the ISO/IEC 25012 standard, it is important to recognize that these characteristics do not depend on specific systems. This technology framework independence makes them particularly valuable as a foundation for policy development. I recommend that by applying these standards, universally applicable policies can be created, both at the individual dataset level and across entire data repositories.

Developing a policy based on ISO/IEC 25012 characteristics facilitates broad compliance and adaptability. This policy can effectively govern data quality management irrespective of the underlying technology used for data storage or processing. This ensures that data quality policies remain effective, even as technology evolves.

In the next section of this thesis, we discuss the findings from Chapter 4 and examine how a policy could improve upon and resolve issues that arose. Table 30 reviews these issues below.

Table 30

Review of Issues

Issue	Description
Metadata Discrepancies and Provenance Issues	In Case Investigation 1, a discrepancy was observed between the metadata update dates Data.gov and Google Dataset Search.

Issue	Description
Accessibility and Server Error Issues	Case Investigation 2 highlighted a significant accessibility issue when attempting to download the dataset resulted in a server error.
Corrupt URLs and Download Errors	In Case Investigation 3, error related to corrupt URLs and failed downloads were documented.
Lower FAIR compliance and Data curation Scores for Google Dataset Search	Lower scores compared to Data.gov.
Lack of Comprehensive Data curation	Across various case investigations, data curation played a limited role in the FAIR compliance of datasets.
FAIR Principles Compliance Issues	Overarching issues with full FAIR compliance in areas of Accessibility, Interoperability, and Reusability.
Validation Issues	Data platforms that complied with FAIR compliance but not data curation.

Application of Policy

As noted in Table 30 – Review of Issues, the first concerns are Metadata Discrepancies and Provenance Issues. A policy based on ISO 25012 would emphasize the importance of accuracy and provenance in metadata. It would require rigorous metadata documentation and update processes. The ISO 25012 inherent characteristics of Accuracy and Credibility would be employed, and by enforcing synchronization between platforms, it would improve dataset reliability and traceability.

The second issue, Accessibility and Server Error Issues, can be addressed using the ISO 25012 characteristics of Accessibility. This characteristic requires that data should be easily and

readily available. Accessibility could be ensured with regular testing and quick-resolution mechanisms for server errors. As such, datasets will always be accessible per FAIR Principles.

A policy using the ISO 25012 characteristics of Accessibility and Reliability could also address the issue of Corrupt URLs and Download Errors. It could be strengthened by ensuring that URLs are regularly checked and maintained. Further, robust systems must be in place to handle download requests without failures.

Lower FAIR compliance and data curation scores resulted when metadata was not applied evenly between Data.gov and Google Dataset Search. The inherent data quality characteristics of Completeness and Compliance in an ISO 25012 policy could resolve this issue by improving metadata completeness. It would also encourage development of a more structured and comprehensive metadata schema. Enforcing these higher standards would improve the overall FAIR compliance scores and the effectiveness of data curation practices.

The policy characteristics of Compliance and Consistency can help resolve the Lack of Comprehensive Data Curation issue. Consistency ensures uniformity is maintained across datasets and platforms—a deficiency highlighted by the vast differences between Google Dataset Search and Data.gov for the same datasets. The ISO 25012-based policy would help embed data curation into the lifecycle of datasets, improving overall utility and integrity.

Another compliance issue that could be resolved concerns missing metadata elements for Accessibility, Interoperability, and Reusability. The Issue with FAIR Principles Compliance identified in the Findings section could be addressed by ISO 25012's inherent characteristics of Completeness, Consistency, and Credibility. These characteristics ensure that metadata is complete (covering all necessary aspects of the data), consistent (uniform across different

datasets), and credible (reliable and trustworthy). This policy focuses on improving metadata quality to support all aspects of FAIR compliance.

Validation issues are the final aspect an ISO 25012-based policy would address. The policy would prioritize the characteristics of accuracy—ensuring data precisely mirrors the real-world context or source from which it originates—and reliability, which guarantees that data can be dependably used for decision-making. To uphold these standards, the policy mandates the use of regular validation processes designed to test and confirm data accuracy and reliability. Such validation is vital for practical data usage and addresses potential gaps in broader FAIR compliance or data curation assessments that might overlook these critical details.

ISO 25012 Policy Implementation Recommendations

Implementing a policy based on the ISO 25012 standard could effectively resolve many challenges identified through case investigations discussed in this thesis. This would lead to enhanced data management practices that adhere to the principles of FAIR compliance and elevate overall data quality. The insights gained from establishing data quality standards according to the ISO/IEC 25012 framework could lay a strong foundation for refining data management practices, supporting the objectives of FAIR compliance by ensuring data is findable, accessible, interoperable, and reusable.

Several strategic approaches can be employed to promote the adoption and implementation of a policy based on ISO 25012 standards. One method is to link compliance through incentives, such as grants and funding. Historical examples from major funding agencies like NIH, NSF, and the European Commission illustrate the effectiveness of such approaches. Providing financial incentives can motivate data repository organizations to adopt ISO 25012 standards in their data management practices. Determination of the funding agency that provides

these financial incentives would most likely be based on the research domain. Additionally, establishing rewards or recognition programs can honor and incentivize researchers and repository organizations that excel in implementing and maintaining high data-quality standards.

The NIH mandates that grant applicants submit DMPs outlining how data will be managed and shared. This aligns with best practices like those in ISO 25012. Compliance with these standards is essential for receiving funding. Additionally, the NSF also requires detailed DMPs as part of the grant application process. These plans must address data quality, preservation, and sharing, which also aligns with ISO 25012 (and similar) principles. The alignment matrix connects the FAIR Principles framework with ISO 25012 Data Quality standards. This connection offers a practical guide for improving RDM processes by integrating these two important concepts into policy. Researchers and data managers can use this matrix to enhance data quality and FAIR compliance simultaneously.

In critical sectors such as healthcare, finance, and public services, integrating an ISO 25012 based policy into mandatory compliance requirements for data management can ensure widespread adoption. For example, in the healthcare sector, non-compliance with such standards could lead to significant penalties imposed by authoritative bodies like the Food and Drug Administration and the U.S. Department of Health and Human Services. In the finance sector, the Securities and Exchange Commission could enforce penalties on organizations that fail to adhere to the ISO 25012-based policies. Similarly, in public services, the National Institute of Standards and Technology could impose fines on organizations that neglect to follow the prescribed policy. Such regulatory measures ensure that the adoption of ISO 25012 standards is not only encouraged but also enforced, thereby enhancing data management practices across these industries.

Open Research FAIR Compliance Recommendation

Several scholars have espoused that the best way to improve open research dataset FAIR compliance is to create metadata standardization—thus enabling researchers to find data they require and be able to access it (Contaxis et al., 2022; Řezník et al., 2022; Rousidis et al., 2015). This thesis shows that while metadata standardization is important, it does not necessarily lead to improvements in open research dataset FAIR compliance. In fact, data curation processes also must be included and followed. A policy that implements standards and provides a roadmap for metadata guidance will ensure successful FAIR compliance and data curation.

Dataset Categorization Recommendation

Establishment of a centralized location where researchers can find datasets remains a worthwhile objective. However, as with Google Dataset Search—and its more than 25 million indexed datasets—limited search options can result in inefficiency and a surplus of unrelated result sets. Populating a metadata field with classification information can provide much better results. The following classification schemes can help enhance search precision:

Categorization by Discipline: Datasets can be classified by broad definitions such as life sciences, engineering, social sciences, etc. This would allow researchers to start their search within a relevant field, reducing the likelihood of encountering unrelated datasets.

Categorization by Usage or Application: Data could also be categorized based on its application areas, such as machine learning, statistical analysis, or educational purposes. This kind of categorization aligns datasets with potential user intentions, facilitating a more targeted search experience.

Hierarchical Categorization: Establishing categories within categories (e.g., under “Life Sciences” having subcategories like “Genetics,” “Neuroscience,” “Bioinformatics”) allows

for a more granular search experience. Users can drill down through layers of categorization to filter their searches.

Dynamic Filtering Options: With robust categorization, platforms can offer dynamic filtering options that adjust available categories based on previous selections. (Think of generative AI.)

Cross-referencing Between Categories: Categorization allows for the possibility of cross-referencing datasets that fall into multiple categories, therefore improving their visibility and usage. Researchers would be able to discover datasets they may not have initially considered but later find relevant.

Further research into implementing categorization strategies can transform the utility of data repositories, turning them from overwhelming stores of information into finely tuned resources that serve the needs of the research community. By populating metadata fields with detailed classification information, platforms can improve the discoverability and accessibility of datasets, supporting more effective research and innovation.

Semantic Keyword Search Recommendation

Semantic keyword search enhances findability by understanding context, relationships, and intent behind search queries, leading to more accurate, relevant, and personalized search results. This technology empowers users to discover information efficiently, even in complex and diverse datasets or across various domains, ultimately improving the overall search experience.

Several areas of semantic keyword search should be examined, regarding fit for purpose. One is Conceptual Understanding, which is achieved through a combination of natural language processing, word embeddings, ontologies, knowledge graphs, contextual analysis, named-entity

recognition, and machine-learning techniques. These technologies work together to help the search engine grasp the meaning and intent behind user queries and deliver more relevant and contextually appropriate search results (Ahmad Khan & Kumar Malik, 2018; Zhong et al., 2002).

Additionally, Natural Language Processing algorithms analyze and interpret human language, enabling the search engine to process and understand natural language queries. NLP algorithms can handle sentence structure, syntax, grammar, and ambiguities inherent in human language. By parsing and understanding input queries, the search engine gains insights into users' intentions, and the concepts they seek.

Ontologies and knowledge graphs can model relationships between concepts and entities. Ontologies are well-defined and structured ways to represent knowledge. Common languages for ontology representation include Resource Description Framework (RDF) and Web Ontology Language (WEB). Because ontologies explicitly define the concepts, properties, and relationships between entities within a domain of interest, they ensure consistency and eliminate ambiguity.

Word Embeddings are an NLP technique that represents words in a continuous vector space, where similar words are located close to each other. These embeddings capture semantic relationships between words based on their co-occurrence patterns in large datasets. When a user enters a search query, the search engine can map words in the query to their corresponding word embeddings. The search engine is then able to understand the conceptual context of the query.

Conclusion

Notably, this thesis illuminated the vulnerabilities of heuristic-based processes, exposing how metadata data can be distorted due to factors such as human error and inadequate control mechanisms during metadata input. Moreover, it underscored the complexity of data governance,

including a lack of data curation standards. It did so by highlighting the shortage of effective methods designed to rectify erroneous data within the data repository, ultimately placing the burden of data verification on researchers.

As we venture into the future of FAIR compliance research, the author of this thesis believes minimizing human intervention during data submission, enforcing metadata standards, and investigating the impacts of dataset categorization on search efficiency should be central to these endeavors. Implementing a policy based on the ISO/IEC 25012 data quality standards could significantly address the issues explored in this thesis.

This study reveals challenges in improving the FAIR compliance of open research datasets, but also provides actionable recommendations for future research. The author proposes integrating FAIR Principles with ISO 25012 data quality standards, offering a practical approach to enhance RDM and FAIR compliance in scientific research.

References

- Ahmad Khan, A., & Kumar Malik, S. (2018, January). Semantic search revisited. In *2018 8th international conference on cloud computing, data science & engineering (confluence)*, (pp. 14–15) IEEE. <https://doi.org/10.1109/CONFLUENCE.2018.8442792>
- Ailamaki, A., Kantere, V., & Dash, D. (2010). Managing scientific data. *Communications of the ACM*, 53(6), 68–78. <https://doi.org/10.1145/1743546.1743568>
- Allen, R., & Hartland, D. (2018). *FAIR in practice—Jisc report on the findable accessible interoperable and reuseable data principles*. Zenodo. <https://doi.org/10.5281/zenodo.1245568>
- Amorim, R. C., Castro, J. A., da Silva, J. R., & Ribeiro, C. (2015). A comparative study of platforms for research data management: interoperability, metadata capabilities and integration potential. In A. Rocha, A. M. Correia, S. Costanzo, & L. P. Reis (Eds.), *New Contributions in Information Systems and Technologies* (pp. 101–111). Springer International Publishing. https://doi.org/10.1007/978-3-319-16486-1_10
- Anderson, C. (2006). The Long Tail: Why the Future of Business Is Selling Less of More. *Hyperion google schola*, 3, 33-50.
- Andrikopoulou, A., Rowley, J., & Walton, G. (2022). Research data management (RDM) and the evolving identity of academic libraries and librarians: A literature review. *New Review of Academic Librarianship*, 28(4), 349-365. <https://doi.org/10.1080/13614533.2021.1964549>
- ARDC. (2024, July 10). *Australian research data commons* <https://ardc.edu.au/>

- Arend, D., Psaroudakis, D., Memon, J. A., Rey-Mazón, E., Schöler, D., Szymanski, J. J., Scholz, U., Junker, A., & Lange, M. (2022). From data to knowledge – big data needs stewardship, a plant phenomics perspective. *The Plant Journal*, *111*(2), 335–347.
<https://doi.org/10.1111/tpj.15804>
- Arpin, S. M., & Kambesis, P. N. (2022). Exploring best practices in data management: examples from cave and karst research and resource management. *Carbonates and Evaporites*, *37*(3), 53. <https://doi.org/10.1007/s13146-022-00772-7>
- Arrow, K. J. (1969). The organization of economic activity: issues pertinent to the choice of market versus nonmarket allocation. *The Analysis and Evaluation of Public Expenditure: The PPB System*, *1*, 59–73.
- Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhler, P., & Wouters, P. (2004). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, *3*, 135–152.
<https://doi.org/10.2481/dsj.3.135>
- Assante, M., Candela, L., Castelli, D., & Tani, A. (2016). Are scientific data repositories coping with research data publishing? *Data Science Journal*, *15*(0), Article 0.
<https://doi.org/10.5334/dsj-2016-006>
- Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399–418.
<https://doi.org/10.1016/j.giq.2015.07.006>
- Ayris, P., & Ignat, T. (2018). Defining the role of libraries in the open science landscape: a reflection on current european practice. *Open Information Science*, *2*(1), 1–22.
<https://doi.org/10.1515/opis-2018-0001>

- Azeroual, O. (2020). Treatment of bad big data in research data management (RDM) systems. *Big Data and Cognitive Computing*, 4(4), Article 4. <https://doi.org/10.3390/bdcc4040029>
- Azeroual, O., Schöpfel, J., Pölönen, J., & Nikiforova, A. (2022). Putting FAIR principles in the context of research information: FAIRness for CRIS and CRIS for FAIRness. *14th International Conference on Knowledge Management and Information Systems (KMIS2022)*, 63–71. <https://doi.org/10.5220/0011548700003335>
- Ball, A. (2012). *Review of data management lifecycle models*. 15.
- Barsness, S., Cummins, J., Fernandez, M. V., James, A. M., Pierce Farrier, K., Pringle, J., Carroll, S. R., Taitingfong, R., & Wieker, A. (2023). *CARE Data Principles, Indigenous data, Data related to Indigenous Peoples and Interest*. <https://hdl.handle.net/11299/256919>
- Bartling, S., & Friesike, S. (Eds.). (2014). *Opening Science: one term, five schools of thought* (pp. 17-47). Springer International Publishing. <https://doi.org/10.1007/978-3-319-00026-8>
- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 41(3), 1-52. <https://doi.org/10.1145/1541880.1541883>
- Beagrie, N. (2006). Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation*, 1, 3–16. <https://doi.org/10.2218/ijdc.v1i1.2>
- Bertagnolli, M. M., Sartor, O., Chabner, B. A., Rothenberg, M. L., Khozin, S., Hugh-Jones, C., Reese, D. M., & Murphy, M. J. (2017). Advantages of a truly open-access data-sharing model. *The New England Journal of Medicine*, 376(12), 1178–1181. <https://doi.org/10.1056/NEJMs1702054>

- Bloemers, M., & Montesanti, A. (2020). The FAIR funding model: providing a framework for research funders to drive the transition toward FAIR data management and stewardship practices. *Data Intelligence*, 2(1–2), 171–180. https://doi.org/10.1162/dint_a_00039
- Bonino da Silva Santos, L. O., Wilkinson, M., Kuzniar, A., Kaliyaperumal, R., Thompson, M., Dumontier, M., & Burger, K. (2016). FAIR data points supporting big data interoperability. *Enterprise Interoperability in the Digitized and Networked Factory of the Future*. ISTE, London, 270–279.
- Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next steps for citizen science. *Science*, 343(6178), 1436–1437. <https://doi.org/10.1126/science.1251554>
- Borgesius, F. Z., Gray, J., & van Eechoud, M. (2015). Open data, privacy, and fair information principles: towards a balancing framework. *Berkeley Technology Law Journal*, 30(3), 2073–2131.
- Borghi, J., Abrams, S., Lowenberg, D., Simms, S., & Chodacki, J. (2018). Support your data: a research data management guide for researchers. *Research Ideas and Outcomes*, 4, e26439. <https://doi.org/10.3897/rio.4.e26439>
- Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059–1078. <https://doi.org/10.1002/asi.22634>
- Borgman, C. L. (2013). *Big data, little data, no data: Scholarship in the networked world*. MIT press. <https://escholarship.org/uc/item/38v6n99v>
- Borgman, C. L. (2021). Big data, little data, or no data? Scholarship, Stewardship, and Humanities Research. <https://escholarship.org/uc/item/5pt0n14g>

- Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, 70(8), 888–904. <https://doi.org/10.1002/asi.24172>
- Boyd, C. (2022). Data as assemblage. *Journal of Documentation*, 78(6), 1338–1352. <https://doi.org/10.1108/JD-08-2021-0159>
- Brickley, D., Burgess, M., & Noy, N. (2019). Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. *The World Wide Web Conference*, (pp.1365–1375). <https://doi.org/10.1145/3308558.3313685>
- Broeck, J. V. den, Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLOS Medicine*, 2(10), e267. <https://doi.org/10.1371/journal.pmed.0020267>
- Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician*, 72(1), 2–10. <https://doi.org/10.1080/00031305.2017.1375989>
- Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science*, 42(5), 351–360. [https://doi.org/10.1002/\(SICI\)1097-4571\(199106\)42:5<351::AID-ASI5>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3)
- Burgess, M. (2020, March 20). What is GDPR? The summary guide to GDPR compliance in the UK. *Wired UK*. <https://www.wired.co.uk/article/what-is-gdpr-uk-eu-legislation-compliance-summary-fines-2018>
- Burton, A., & Treloar, A. (2009). Designing for discovery and re-use: the ‘ANDS data sharing verbs’ approach to service decomposition. *International Journal of Digital Curation*, 4(3), Article 3. <https://doi.org/10.2218/ijdc.v4i3.124>

- Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining data information literacy needs: a study of students and research faculty. *Portal: Libraries and the Academy*, 11(2), 629–657.
- Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for Indigenous data governance. *Data Science Journal*, 19, 43–43. <https://doi.org/10.5334/dsj-2020-043>
- Carroll, S. R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR principles for Indigenous data futures. *Scientific Data*, 8(1), 108. <https://doi.org/10.1038/s41597-021-00892-0>
- Caulley, D. N. (2007). Conducting research literature reviews: from the internet to paper. *Qualitative Research Journal*, 7(2), 103–105.
- Chancellor, V. (2020). Patron (Col.) Prof. S.S. Sarangdevot. *Arthavati*, 7(1 & 2), 25.
- Charalabidis, Y., Alexopoulos, C., Lampoltshammer, T., Zuiderwijk, A., Janssen, M., & Ferro, E. (2018). The world of open data: concepts, methods, tools and experiences. *Public Administration and Information Technology*, 28, 1–194. https://doi.org/10.1007/978-3-319-90850-2_1
- Chignard, S. (2013, March 29). A brief history of open data. *ParisTech Review*. <http://www.paristechreview.com/2013/03/29/brief-history-open-data/>
- Coase, R. H. (1993). The nature of the firm (1937). *Williamson, OE; Winter, SG*. <http://econdse.org/wp-content/uploads/2014/09/firm-coase.pdf>

- Coglianese, C. (2009). The transparency president? The Obama administration and open government. *Governance*, 22(4), 529–544. <https://doi.org/10.1111/j.1468-0491.2009.01451.x>
- Contaxis, N., Clark, J., Dellureficio, A., Gonzales, S., Mannheimer, S., Oxley, P. R., Ratajeski, M. A., Surkis, A., Yarnell, A. M., Yee, M., & Holmes, K. (2022). Ten simple rules for improving research data discovery. *PLOS Computational Biology*, 18(2), e1009768. <https://doi.org/10.1371/journal.pcbi.1009768>
- Conway, P. (2010, January). Preservation in the age of Google: digitization, digital preservation, and dilemmas. *The Library Quarterly: Vol 80, No 1*. <https://www-journals-uchicago-edu.libezproxy2.syr.edu/doi/full/10.1086/648463>
- Corti, L., Woollard, M., Bishop, L., & Van den Eynden, V. (2019). Managing and sharing research data: A Guide to Good Practice. 1–368.
- Creswell, J., & Poth, C. (2024, June 13). Qualitative inquiry and research design. *SAGE Publications Inc*. <https://us.sagepub.com/en-us/nam/qualitative-inquiry-and-research-design/book266033>
- National Institutes of Health. (2023, January). Data management & sharing policy overview | data sharing. <https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policies/data-management-and-sharing-policy-overview>
- Davies, T., & Perini, F. (2016). Researching the emerging impacts of open data: revisiting the ODDC conceptual framework. *The Journal of Community Informatics*, 12(2), Article 2. <https://doi.org/10.15353/joci.v12i2.3246>

- Devaraju, A., & Berkovsky, S. (2018). A hybrid recommendation approach for open research datasets. *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 207–211. <https://doi.org/10.1145/3209219.3209250>
- Devine, D. (2024, March 18). Examining dataset FAIR compliance in the research data management lifecycle.
- Directorate-General for Communications Networks, C. and T., International, I., con.terra, Commission, E., Fokus, F., Southampton, U. of, Media, D.-G. for the I. S. and, Time.lex, Institute, O. D., Sogeti, & Consulting, C. (n.d.). Creating value through open data: study on the impact of re-use of public data resources | Policy commons. Retrieved June 21, 2024, from <https://policycommons-net.libezproxy2.syr.edu/artifacts/262863/creating-value-through-open-data/1045680/>
- Directorate-General for Research and Innovation (European Commission). (2018). Cost-benefit analysis for FAIR research data: cost of not having FAIR research data. *Publications Office of the European Union*. <https://data.europa.eu/doi/10.2777/02999>
- Donaldson, D. R., & Koepke, J. W. (2022). A focus groups study on data sharing and research data management. *Scientific Data*, 9(1), Article 1. <https://doi.org/10.1038/s41597-022-01428-w>
- Douglass, K., Allard, S., Tenopir, C., Wu, L., & Frame, M. (2014). Managing scientific data as public assets: data sharing practices and policies among full-time government employees. *Journal of the Association for Information Science and Technology*, 65(2), 251–262. <https://doi.org/10.1002/asi.22988>

- Duke, J. E. (2022). U.S. Department of the Interior: sharing FAIR data fairly [*Ph.D., University of Arkansas at Little Rock*].
<https://www.proquest.com/pqdtglobal/docview/2771680420/abstract/B532C6148CEE40F5PQ/2>
- Dunning, A., Smaele, M. de, & Böhmer, J. (2017). Are the FAIR data principles fair?
International Journal of Digital Curation, 12(2), Article 2.
<https://doi.org/10.2218/ijdc.v12i2.567>
- Dutch Techcentre for Life Sciences. (2024), What is FAIR data stewardship?
<https://www.dtls.nl/fair-data/data-stewardship/>
- Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata principles and practicalities. *D-Lib Magazine*, 8(4). <https://doi.org/10.1045/april2002-weibel>
- Eaker, C. (2016). What could possibly go wrong? The impact of poor data management. 27. *Elsevier*. (2022, October 6). *Research Data*. <https://www.elsevier.com/about/policies/research-data>
- European Commission. (2024, June 10). European legislation on open data - shaping Europe's digital future.. <https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data>
- European Union. (2019). Directive (EU) 2019/1024 of the European parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (*Recast*), *EP, CONSIL, 172 OJ L*/ <http://data.europa.eu/eli/dir/2019/1024/oj/eng>
- Fecher, B., Bartling, S., & Friesike, S. (2014). Opening science: the evolving guide on how the internet is changing research, collaboration and scholarly publishing. *Impact of Social Sciences Blog*. <https://eprints.lse.ac.uk/71034/>

- Fecher, B., & Friesike, S. (2014). Open Science: One Term, Five Schools of Thought. *Springer International Publishing*.
https://library.oapen.org/bitstream/handle/20.500.12657/28008/1/2014_Book_OpeningScience.pdf#page=24
- Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PloS One*, *10*(2), e0118053.
- Fink, A. (2019, February 1). Conducting research literature reviews. *SAGE Publications Inc*.
<https://us.sagepub.com/en-us/nam/conducting-research-literature-reviews/book259191>
- Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, *1*(3923), 554.
- Foundation, N.-N. S. (2017). Dissemination and sharing of research results.
- Fry, J., Lockyer, S., Oppenheim, C., Houghton, J., & Rasmussen, B. (2009). Identifying benefits arising from the curation and open sharing of research data. *Produced by UK Higher Education and research institutes*.
- Garnett, A., Leahey, A., Savard, D., Towell, B., & Wilson, L. (2017). Open metadata for research data discovery in Canada. *Journal of Library Metadata*, *17*(3–4), 201–217.
<https://doi.org/10.1080/19386389.2018.1443698>
- Gilliland, A. J. (2008). Setting the stage. *Introduction to Metadata*, *2*(1–19), 7.
- Gonzalez, A., & Peres-Neto, P. (2015). Data curation: act to staunch loss of research data. *Nature*, *520*, 436. <https://doi.org/10.1038/520436c>
- Gregory, K., Khalsa, S. J., Michener, W. K., Psomopoulos, F. E., Waard, A. de, & Wu, M. (2018). Eleven quick tips for finding research data. *PLOS Computational Biology*, *14*(4), e1006038. <https://doi.org/10.1371/journal.pcbi.1006038>

- Gualo, F., Rodriguez, M., Verdugo, J., Caballero, I., & Piattini, M. (2021). Data quality certification using ISO/IEC 25012: industrial experiences. *Journal of Systems and Software*, 176, 110938. <https://doi.org/10.1016/j.jss.2021.110938>
- Guibault, L., & Wiebe, A. (Eds.). (2013). Safe to be open: study on the protection of research data and recommendations for access and usage. *Universitätsverlag Göttingen*. <https://doi.org/10.17875/gup2013-160>
- Halevy, A., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., & Whang, S. E. (2016). Goods: organizing Google's datasets. *Proceedings of the 2016 International Conference on Management of Data*, 795–806. <https://doi.org/10.1145/2882903.2903730>
- Hanson, B., Sugden, A., & Alberts, B. (2011). Making data maximally available. *Science*, 331(6018), 649.
- Harrow, I., & Liener, T. (2021). Implementing the FAIR data principles is now a critical endeavour. *European Pharmaceutical Review*, 26(3), 56–57.
- Hauschke, C., Nazarovets, S., Altemeier, F., & Kaliuzhna, N. (2021). Roadmap to FAIR research information in open infrastructures. *Journal of Library Metadata*, 21(1–2), 45–61. <https://doi.org/10.1080/19386389.2021.1999156>
- Hjørland, B., & Wilson, P. (1997). Information seeking and subject representation: an activity-theoretical approach to information science. In *New Directions in Information Management*.
- Hodson, S. (2011). Managing and sharing data: best practice for researchers.
- Humphreys, S. (2018). Data: the given. In J. Hohmann & D. Joyce (Eds.), *International Law's Objects* (p. 0). Oxford University Press. <https://doi.org/10.1093/oso/9780198798200.003.0016>

- Hunter, J., & Iannella, R. (1998). The application of metadata standards to video indexing. In C. Nikolaou & C. Stephanidis (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 135–156). Springer. https://doi.org/10.1007/3-540-49653-X_9
- Imperial College London. (2022, October 8). What is research data? (<https://www.imperial.ac.uk/research-and-innovation/support-for-staff/scholarly-communication/research-data-management/introduction-to-research-data-management/what-is-research-data/>)
- ISO - Standards. (n.d.). ISO. Retrieved April 27, 2024, from <https://www.iso.org/standards.html>
- Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., ... Schultes, E. (2020). FAIR principles: interpretations and implementation considerations. *Data Intelligence*, 2(1–2), 10–29. https://doi.org/10.1162/dint_r_00024
- Jansen, B. J., & Pooch, U. (2001). A Review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52(3), 235–246. [https://doi.org/10.1002/1097-4571\(2000\)9999:9999<::AID-ASI1607>3.0.CO;2-F](https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1607>3.0.CO;2-F)
- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4), 258–268. <https://doi.org/10.1080/10580530.2012.716740>
- Jati, P. H. P., Lin, Y., Nodehi, S., Cahyono, D. B., & van Reisen, M. (2022). FAIR versus open data: a comparison of objectives and principles. *Data Intelligence*, 4(4), 867–881. https://doi.org/10.1162/dint_a_00176

- Juty, N., Wimalaratne, S. M., Soiland-Reyes, S., Kunze, J., Goble, C. A., & Clark, T. (2020). Unique, persistent, resolvable: identifiers as the foundation of FAIR. *Data Intelligence*, 2(1–2), 30–39. https://doi.org/10.1162/dint_a_00025
- Kagolovsky, Y., & Moehr, J. R. (2003). Current status of the evaluation of information retrieval. *Journal of Medical Systems*, 27(5), 409–424. <https://doi.org/10.1023/A:1025603704680>
- Kahn, M. G., Brown, J. S., Chun, A. T., Davidson, B. N., Meeker, D., Ryan, P. B., Schilling, L. M., Weiskopf, N. G., Williams, A. E., & Zozus, M. N. (2015). Transparent reporting of data quality in distributed data networks. *eGEMs*, 3(1), 1052. <https://doi.org/10.13063/2327-9214.1052>
- Kapiszewski, D., & Karcher, S. (2020). Making research data accessible. In C. Elman, J. Gerring, & J. Mahoney (Eds.), *The Production of Knowledge* (1st ed., pp. 197–220). Cambridge University Press. <https://doi.org/10.1017/9781108762519.008>
- Keene, S. (2002). Preserving digital materials: confronting tomorrow’s problems today. *The Conservator*, 26(1), 93–99. <https://doi.org/10.1080/01410096.2002.9995181>
- Koltay, T. (2017, March 1). Data literacy for researchers and data librarians. <https://doi.org/10.1177/0961000615616450>
- Labadie, C., Legner, C., Eurich, M., & Fadler, M. (2020). FAIR enough? Enhancing the usage of enterprise data with data catalogs. *2020 IEEE 22nd Conference on Business Informatics (CBI)*, 1, 201–210. <https://doi.org/10.1109/CBI49978.2020.00029>
- Langer, A., & Bilz, E. (2019). Analysis of current RDM applications for the interdisciplinary *Publication of Research Data*. 12.
- Lawal, I. (2010). Ensuring the integrity, accessibility, and stewardship of research data in the digital age (review). *Portal: Libraries and the Academy*, 10(3), 365–366.

- Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST principles for digital repositories. *Scientific Data*, 7(1), Article 1.
<https://doi.org/10.1038/s41597-020-0486-7>
- Liu, F., Xiao, B., Lim, E. T. K., & Tan, C.-W. (2016). Is my effort worth it? Investigating the dual effects of search cost on search utility. *AIS Electronic Library (AISel)*.
- Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation (Vol. 67).
- Lynch, C. (2008). How do your data grow? *Nature*, 455(7209), Article 7209.
<https://doi.org/10.1038/455028a>
- Mahdi, M. N., Ahmad, A. R., Ismail, R., Subhi, M. A., Abdulrazzaq, M. M., & Qassim, Q. S. (2020). Information overload: the effects of large amounts of information. *2020 1st. Information Technology To Enhance e-Learning and Other Application (IT-ELA)*, 154–159. <https://doi.org/10.1109/IT-ELA50150.2020.9253082>
- Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Doshi, E. A. (2013). Open data: unlocking innovation and performance with liquid information. *McKinsey Global Institute*, 21, 116.
- Mayernik, M. S. (2016). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, 67(4), 973–993.
<https://doi.org/10.1002/asi.23425>
- McKiernan, E., Bourne, P., Brown, C. T., & Buck, S. (2016, July 7). *Point of view: how open science helps researchers succeed* / *eLife*. <https://elifesciences.org/articles/16800>

- McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., & Sansone, S.-A. (2016). BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database*, 2016, baw075.
- Mellinger, C. D., & Hanson, T. A. (2020). Methodological considerations for survey research: validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 19. <https://doi.org/10.52034/lanstts.v19i0.549>
- Michener, W. K. (2015). Ecological data sharing. *Ecological Informatics*, 29, 33–44. <https://doi.org/10.1016/j.ecoinf.2015.06.010>
- Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, 27(2), 85–93. <https://doi.org/10.1016/j.tree.2011.11.016>
- Miyakawa, T. (2020). No raw data, no science: another possible source of the reproducibility crisis. *Molecular Brain*, 13(1), 24. <https://doi.org/10.1186/s13041-020-0552-2>
- Molloy, J. C. (2011). The open knowledge foundation: open data means better science. *PLOS Biology*, 9(12), e1001195. <https://doi.org/10.1371/journal.pbio.1001195>
- Mons, B. (2018). *Data stewardship for open science: implementing FAIR principles*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781315380711>
- Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European open science cloud. *Information Services & Use*, 37(1), 49–56. <https://doi.org/10.3233/ISU-170824>
- National Academies of Sciences, Engineering, and Medicine. (2018). *Open science by design: realizing a vision for 21st century research*. <https://doi.org/10.17226/25116>.

- Nosek, B. A., Alter, G., & Banks, G. C. (2015, August 26). Promoting an open research culture. *Science*. <https://www-science-org.libezproxy2.syr.edu/doi/full/10.1126/science.aab2374>
- National Institutes of Health (2023, January). *Data management & sharing policy overview / data sharing*. <https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policies/data-management-and-sharing-policy-overview>
- National Science Foundation (2024, July 18). <https://www.nsf.gov/>
- Open Knowledge (2022) What is open? <https://okfn.org>
- Owan, V. J., & Bassey, B. A. (2019). Data management practices in educational research (*SSRN Scholarly Paper* 3516191). <https://papers.ssrn.com/abstract=3516191>
- Paic, A. (2021). *Open science—enabling discovery in the digital age*. <https://doi.org/10.1787/81a9dcf0-en>
- Palmer, C. L. (2009). Scholarly information practices in the online environment: themes from the literature and implications for library service development. *OCLC*.
- Pampel, H., & Dallmeier-Tiessen, S. (2014). Open research data: from vision to practice. *Opening Science: The Evolving Guide on How the Internet Is Changing Research, Collaboration and Scholarly Publishing*, 213–224.
- Parmiggiani, E., & Grisot, M. (2020). Data curation as governance practice. 32. <https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2673540>
- Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). *On the reuse of scientific data*. <https://doi.org/10.5334/dsj-2017-008>
- Patel, D. (2016). Research data management: a conceptual framework. *Library Review*, 65(4/5), 226–241. <https://doi.org/10.1108/LR-01-2016-0001>
- Pearson, K. (1957). *The grammar of science*. Рипол Классик.

- Peng, G., Lacagnina, C., Ivánová, I., Downs, R. R., Ramapriyan, H., Ganske, A., Jones, D., Bastin, L., Wyborn, L., Bastrakova, I., Wu, M., Shie, C.-L., Moroni, D. F., Larnicol, G., Wei, Y., Ritchey, N., Champion, S., Hou, C.-Y., Habermann, T., ... Roux, J. le. (2021). *International community guidelines for sharing and reusing quality information of individual earth science datasets*. OSF. <https://doi.org/10.31219/osf.io/xsu4p>
- Peng, G., Privette, J. L., Kearns, E. J., Ritchey, N. A., & Ansari, S. (2015). A unified Framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, 13, 231–253. <https://doi.org/10.2481/dsj.14-049>
- Peng, R. (2015). The reproducibility crisis in science: a statistical counterattack. *Significance*, 12(3), 30–32. <https://doi.org/10.1111/j.1740-9713.2015.00827.x>
- Pinfield, S., Cox, A. M., & Smith, J. (2014). Research data management and libraries: relationships, activities, drivers and influences. *PLOS ONE*, 9(12), e114734. <https://doi.org/10.1371/journal.pone.0114734>
- Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Commun. ACM*, 45(4), 211–218. <https://doi.org/10.1145/505248.506010>
- Piwowar, H. A., Priem, J., Lariviere, V., & Alperin, J. P. (2018, February 13). *The State of OA: a Large-Scale Analysis of the Prevalence and Impact of Open Access Articles [PeerJ]*. <https://peerj.com/articles/4375/?ref=blog.oa.works>
- Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1, e175. <https://doi.org/10.7717/peerj.175>
- Publications Office of the European Union. (2020). *The Economic Impact of Open Data: Opportunities for Value Creation in Europe*. Publications Office. <https://data.europa.eu/doi/10.2830/63132>

- Qin, J., Crowston, K., & Kirkland, A. (2017). Pursuing best performance in research data management by using the capability maturity model and rubrics. *Journal of eScience Librarianship*, 6(2), e1113. <https://doi.org/10.7191/jeslib.2017.1113>
- Quimbert, E., Jeffery, K., Martens, C., Martin, P., & Zhao, Z. (2020). Data Cataloguing. In Z. Zhao & M. Hellström (Eds.), *Towards Interoperable Research Infrastructures for Environmental and Earth Sciences: A Reference Model Guided Approach for Common Challenges* (pp. 140–161). Springer International Publishing.
https://doi.org/10.1007/978-3-030-52829-4_8
- Radley-Gardner, O., Beale, H., & Zimmermann, R. (Eds.). (2016). *Fundamental Texts On European Private Law* (2nd ed.). Hart Publishing Ltd.
<https://doi.org/10.5040/9781782258674>
- Reinsel, D., Gantz, J., & Rydning, J. (2018). *The Digitization of the World from Edge to Core*.
- Renear, A. H., & Palmer, C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing. *Science*, 325(5942), 828–832. <https://doi.org/10.1126/science.1157784>
- Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, 47(1), 1–4. <https://doi.org/10.1002/meet.14504701240>
- Resnik, D. B., & Shamoo, A. E. (2017). Reproducibility and research integrity. *Accountability in Research*, 24(2), 116–123. <https://doi.org/10.1080/08989621.2016.1257387>
- Řezník, T., Raes, L., Stott, A., De Lathouwer, B., Perego, A., Charvát, K., & Kafka, Š. (2022). Improving the documentation and findability of data services and repositories: a review of (meta)data management approaches. *Computers & Geosciences*, 169, 105194.
<https://doi.org/10.1016/j.cageo.2022.105194>

- Riley, J. (2017). *Understanding Metadata*. 49.
- Rosenthal, D. S. H., Robertson, T. S., Lipkis, T., Reich, V., & Morabito, S. (2005). *Requirements for Digital Preservation Systems: A Bottom-Up Approach* (arXiv:cs/0509018). arXiv. <https://doi.org/10.48550/arXiv.cs/0509018>
- Rousidis, D., Garoufallou, E., Balatsoukas, P., & Sicilia, M.-A. (2015). Evaluation of metadata in research data repositories: the case of the DC subject element. In E. Garoufallou, R. J. Hartley, & P. Gaitanou (Eds.), *Metadata and Semantics Research* (pp. 203–213). Springer International Publishing. https://doi.org/10.1007/978-3-319-24129-6_18
- Ruggles, S. (2018). The importance of data curation. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research* (pp. 303–308). Springer International Publishing. https://doi.org/10.1007/978-3-319-54395-6_39
- Rusbridge, C. (2007, May 2). *Create, Curate, Re-use: The Expanding Life Course of Digital Research Data*. <https://era.ed.ac.uk/handle/1842/1731>
- Scherle, G. J., Hollie C. White, Sarah Carrier, Ryan. (2012). A metadata best practice for a scientific data repository. In *Metadata Best Practices and Guidelines*. Routledge.
- Schöpfel, J., Ferrant, C., André, F., & Fabre, R. (2018). Research data management in the French National Research Center (CNRS). *Data Technologies and Applications*, 52(2), 248–265. <https://doi.org/10.1108/DTA-01-2017-0005>
- Sheridan, H., Dellureficio, A. J., Ratajeski, M. A., Mannheimer, S., & Wheeler, T. R. (2021). Data curation through catalogs: a repository-independent model for data discovery. *Journal of eScience Librarianship*, 10(3), Article 3. <https://doi.org/10.7191/jeslib.2021.1203>

- Simonetta, A., Paoletti, M. C., & Venticinque, A. (2021). *Using the SQuaRE Series as a Guarantee for GDPR Compliance*.
- Škoda, P., Bernhauer, D., Nečaský, M., Klímek, J., & Skopal, T. (2020). Evaluation framework for search methods focused on dataset findability in open data catalogs. *Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services*, 200–209. <https://doi.org/10.1145/3428757.3429973>
- Skopal, T., Klímek, J., & Nečaský, M. (2019). Improving findability of open data beyond data catalogs. *Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services*, 413–417. <https://doi.org/10.1145/3366030.3366095>
- Spicer, R. A. (2018). *Fit for purpose? A metascientific analysis of metabolomics data in public repositories* (Publication No. 34945). [Doctoral dissertation, University of Cambridge]. Apollo. <https://doi.org/10.17863/CAM.34945>
- Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., & Wyborn, L. (2019). Make scientific data FAIR. *Nature*, 570 (7759), 27–29. <https://doi.org/10.1038/d41586-019-01720-7>
- Stevens, H. (2016). Big data, little data, no data: Scholarship in the networked world. by Christine L. Borgman (review). *Technology and Culture*, 57(3), 706–708.
- Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by Journals. *PLOS ONE*, 8(6), e67111. <https://doi.org/10.1371/journal.pone.0067111>
- Suber, P. (2012). *Open access*. MIT Press.

- Surkis, A., & Read, K. (2015). Research data management. *Journal of the Medical Library Association : JMLA*, 103(3), 154–156. <https://doi.org/10.3163/1536-5050.103.3.011>
- Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, Chris. H. J. (2016). The academic, economic and societal impacts of open access: An evidence-based review. *F1000Research*, 5, 632. <https://doi.org/10.12688/f1000research.8460.3>
- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLOS ONE*, 6(6), e21101. <https://doi.org/10.1371/journal.pone.0021101>
- Thorsby, J., Stowers, G. N. L., Wolslegel, K., & Tumbuan, E. (2017). Understanding the content and features of open data portals in American cities. *Government Information Quarterly*, 34(1), 53–61. <https://doi.org/10.1016/j.giq.2016.07.001>
- Tolle, K. M., Tansley, D. S. W., & Hey, A. J. G. (2011). The fourth paradigm: Data-intensive scientific discovery [Point of View]. *Proceedings of the IEEE*, 99(8), 1334–1337. *Proceedings of the IEEE*. <https://doi.org/10.1109/JPROC.2011.2155130>
- Tsueng, G., Cano, M. A. A., Bento, J., Czech, C., Kang, M., Pache, L., Rasmussen, L. V., Savidge, T. C., Starren, J., Wu, Q., Xin, J., Yeaman, M. R., Zhou, X., Su, A. I., Wu, C., Brown, L., Shabman, R. S., & Hughes, L. D. (2023). Developing a standardized but extendable framework to increase the findability of infectious disease datasets. *Scientific Data*, 10(1), Article 1. <https://doi.org/10.1038/s41597-023-01968-9>

- Turner, J. A., Calhoun, V. D., Thompson, P. M., Jahanshad, N., Ching, C. R. K., Thomopoulos, S. I., Verner, E., Strauss, G. P., Ahmed, A. O., Turner, M. D., Basodi, S., Ford, J. M., Mathalon, D. H., Preda, A., Belger, A., Mueller, B. A., Lim, K. O., & van Erp, T. G. M. (2022). ENIGMA + COINSTAC: Improving findability, accessibility, interoperability, and re-usability. *Neuroinformatics*, 20(1), 261–275. <https://doi.org/10.1007/s12021-021-09559-y>
- Ubaldi, B. (2013). *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives* (OECD Working Papers on Public Governance 22; OECD Working Papers on Public Governance, Vol. 22). <https://doi.org/10.1787/5k46bj4f03s7-en>
- van Belle, G., & Ruiter, L. (2014). Data and the law: Beyond the sweat of the brow: Who owns published data? And what is data? *Significance*, 11(2), 28–31. <https://doi.org/10.1111/j.1740-9713.2014.00737.x>
- Verbaan, E., & Cox, A. M. (2014). Occupational sub-cultures, jurisdictional struggle and third space: Theorising professional service responses to research data management. *The Journal of Academic Librarianship*, 40(3), 211–219. <https://doi.org/10.1016/j.acalib.2014.02.008>
- Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88, 428–436. <https://doi.org/10.1016/j.jbusres.2017.12.043>
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*. <https://www.tandfonline.com/doi/abs/10.1080/07421222.1996.11518099>

- Weagley, J., Gelches, E., & Park, J.-R. (2010). Interoperability and metadata quality in digital video repositories: A study of Dublin core. *Journal of Library Metadata*, 10(1), 37–57.
<https://doi.org/10.1080/19386380903546984>
- Weibel, S. L., & Koch, T. (2000). The Dublin core metadata initiative: mission, current activities, and future directions. *D-Lib Magazine*, 6(12).
[https://doi.org/10.1045/december2000-weibel`](https://doi.org/10.1045/december2000-weibel)
- Imperial College London. (2022, October 8). *What is Research Data?*
(<https://www.imperial.ac.uk/research-and-innovation/support-for-staff/scholarly-communication/research-data-management/introduction-to-research-data-management/what-is-research-data/>)
- Whyte, A., & Tedds, J. (2011, September 1). *Making the Case for Research Data Management / DCC*. <https://www.dcc.ac.uk/guidance/briefing-papers/making-case-rdm>
- Wiggins, A., & Wilbanks, J. (2019). Citizen science: The theory and practice of public participation in scientific research. *Journal of Science Communication*, 18(1).
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), Article 1.
<https://doi.org/10.1038/sdata.2016.18>

- Wilkinson, M. D., Verborgh, R., Santos, L. O. B. da S., Clark, T., Swertz, M. A., Kelpin, F. D. L., Gray, A. J. G., Schultes, E. A., Mulligen, E. M. van, Ciccarese, P., Kuzniar, A., Gavai, A., Thompson, M., Kaliyaperumal, R., Bolleman, J. T., & Dumontier, M. (2017). Interoperability and FAIRness through a novel combination of web technologies. *PeerJ Computer Science*, 3, e110. <https://doi.org/10.7717/peerj-cs.110>
- Wolfram, D. (2008). Search characteristics in different types of web-based IR environments: Are they the same? *Information Processing & Management*, 44(3), 1279–1292. <https://doi.org/10.1016/j.ipm.2007.07.010>
- Zhang, B., Pouchard, L. C., Smith, P. M., Gasc, A., & Pijanowski, B. C. (2016). *Data Storage and Sharing for the Long Tail of Science*. 2016 New York Scientific Data Summit, NYSDS 2016 - Proceedings. <https://doi.org/10.1109/NYSDS.2016.7747811>
- Zhang, X.-F. (2021). Application of blockchain technology in data management of university scientific research. *Advances in Intelligent Systems and Computing*, 1195 AISC, 606–613. https://doi.org/10.1007/978-3-030-50399-4_60
- Zhong, J., Zhu, H., Li, J., & Yu, Y. (2002). Conceptual graph matching for semantic search. In U. Priss, D. Corbett, & G. Angelova (Eds.), *Conceptual Structures: Integration and Interfaces* (pp. 92–106). Springer. https://doi.org/10.1007/3-540-45483-7_8

Vita

Author: Denise L. Devine

Contact: dldevine@syr.edu

Education:

Bachelor of Business Administration, University of Memphis, 1993

Master of Science in Applied Data Science, Syracuse University, 2019

Doctor of Professional Studies in Information Management, Syracuse University, 2024

Professional Experience:

Over 30 years in Information Technology. Primarily focused on SQL Server technology including DBA, database developer and data analysis.