12-1-2018

# The Impact of Research Data Sharing and Reuse on Data Citation in STEM Fields

Hyoungjoo Park
*University of Wisconsin-Milwaukee*

# THE IMPACT OF RESEARCH DATA SHARING AND REUSE ON DATA CITATION IN STEM FIELDS

by

Hyoungjoo Park

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Information Studies

at

The University of Wisconsin -Milwaukee

December 2018

# ABSTRACT

THE IMPACT OF RESEARCH DATA SHARING AND REUSE ON DATA CITATION IN STEM FIELDS

by

Hyoungjoo Park

The University of Wisconsin Milwaukee, 2018

Under the Supervision of Dr. Dietmar Wolfram

Despite the open science movement and mandates for the sharing of research data by major funding agencies and influential journals, the citation of data sharing and reuse has not become standard practice in the various science, technology, engineering and mathematics (STEM) fields. Advances in technology have lowered some barriers to data sharing, but it is a socio-technical phenomenon and the impact of the ongoing evolution in scholarly communication practices has yet to be quantified. Furthermore, there is need for a deeper and more nuanced understanding of author self-citation and recitation, the most often cited types of data, disciplinary differences regarding data citation and the extent of interdisciplinarity in data citation.

This study employed a mixed methods approach that combined coding with semi-automatic text-searching techniques in order to assess the impact of data sharing and reuse on data citation in STEM fields. The research considered over 500,000 open research data entities, such as datasets, software and data studies, from over 350 repositories worldwide. I also examined 705 bibliographic publications with a total of 15,261 instances of data sharing, reuse, and citation the data, article, discipline and interdisciplinary levels. More specifically, I measured the phenomenon of data sharing in terms of formal data citation, frequently cited data types, and author self-citation, and I explored recitation at the levels of both data- and bibliography-level, and data reuse practices in bibliographies, associations of disciplines, and interdisciplinary contexts.

The results of this research revealed, to begin with, disciplinary differences with regard to the impact of data sharing and reuse on data citation in STEM fields. This research also yielded the following additional findings regarding the citation of data by STEM researchers; 1) data sharing practices were diverse across disciplines; 2) data sharing has been increasing in recent years; 3) each discipline made use of major digital repositories; 4) these repositories took various forms depending on the discipline; 5) certain data types were more often cited in each discipline, so that the frequency distribution of the data types was highly skewed; 6) author self-citation and recitation followed similar trends at the data and bibliographic levels, but specific practices varied within each discipline; 7) associations between and across data and author self-citation and recitation at the bibliographic level were observed, with the self-citation rate differing significantly among disciplines;8) data reuse in bibliographies was rare yet diverse; 9) informal citation of data sharing and reuse at the bibliographic level was more common in certain fields, with astronomy/physics showing the highest amount (98%) and technology the lowest (69%); 10) within bibliographic publications, the documentation of data sharing and reuse occurred mainly in the main text; 11) publications in certain disciplines, such as chemistry, computing and engineering, did not attract citations from more than one field (i.e., showed no diversity); and, on the other hand,12) publications in other fields attracted a wide range of interdisciplinary data citations.

This dissertation, then, contributes to the understanding of two key areas aspects of the current citation systems. First, the findings have practical implications for individual researchers, decision makers, funding agencies and publishers with regard to giving due credits to those who share their data. Second, this research has methodological implications in terms of reducing the labor required to analyze the full text of associated articles in order to identify evidence of data citation.

To

my parents

and my sister

# TABLE OF CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| DCI | Data Citation Index |
| DOIs | Data Object Identifier |
| OA | Open Access |
| OPR | Open Peer Review |
| OS | Open Science |
| SCI | Science Citation Index |
| STEM | Science Technology Engineering and Mathematics |
| WoS | Web of Science |

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

First and foremost, I offer my deepest appreciation and sincere gratitude to my advisor, Dr. Dietmar Wolfram (chair); I consider myself blessed to have such a caring and inspiring mentor. In weekly meetings, Dr. Wolfram generously shared his knowledge and time and helped me to learn as I proceeded step by step thanks to his stimulating advice, insightful guidance, and warm encouragement throughout this study. I extend my appreciation to the other members of my committee as well, Dr. Richard Smiraglia from Knowledge Organization, Dr. Jacques du Plessis from Open Science, Dr. Margaret Kipp from Open Software, and Dr. Catherine Blake from text searching, who provided helpful comments from their various perspectives. I am also thankful to all of the faculty, staff members, and fellow doctoral students in the School of Information Studies at the University of Wisconsin Milwaukee who rooted for me and served as a major source of inspiration, encouragement, and support. I would further like to acknowledge my professors at Syracuse University, who provided rigorous education and training that developed my abilities to carry out practical research in information management. I also would like to acknowledge my professors at Sungkyunkwan University (SKKU) in South Korea, who encouraged me and helped to build the foundation of my knowledge in library and information science, in particular my supervisor, Dr. Sam Oh, for his caring advice and continuing warm support. My warmest appreciation goes to my parents, whose love, care, and support have been a constant source of strength and motivation for me. My heart-felt gratitude goes to my sister too, who has always been my very best friend and a great colleague who keeps me going. Finally, I thank God for equipping me, through His Holy Spirit, with strength, knowledge, ability, and fortitude, as well as specific wisdom and strength in my academic endeavors.

# Chapter 1 INTRODUCTION

## 1.1. Research Problem and Motivation

Open science is an umbrella term for an approach based on greater access to public research that can affect the entire research cycle and its stakeholders and can be enhanced by information and communication technology (ICT) tools, platforms, networked collaboration, and participation, all of which promote the diffusion of research results. The open science movement supports reuse, reproducibility, and transparency. Open data—data that are freely and openly available to the general public—are widely used in scholarly communication, governmental, and industrial sectors. Aspects of open science that are publicly visible and/or citable include open research data, Open Access (OA) journals (e.g., Public Library of Science ONE or PLOS ONE) that may also employ open peer review (e.g., F1000Research), publicly accessible repositories (e.g., Harvard Dataverse), open source software (e.g., Apache OpenOffice), and various other open scholarship practices. The OA option is associated with higher citation rates in comparison with non-OA articles published in the same journals: OA articles twice as likely to be cited within 4 to 10 months and three times more likely after 10 to 16 months (Eysenbach, 2006) and are more often downloaded from publishers' websites than closed access articles (Davis, 2010). Open science with shared data can have a relatively greater impact (Piwowar, Day, & Fridsma, 2007), increasing reproducibility through data reuse and repurposing research questions and fostering transparency through the validation of research findings. The sustainability of open science is dependent on maximizing data reuse rather than the mere sharing of data in repositories (Curty, 2015) because data reuse promotes data sharing (Niu, 2009). Open science, which includes open access to research data, can help researchers to succeed in these respects.

In this era of big data, open science, and open research data, it is becoming increasingly important to measure the scholarly impact of data sharing on research data citation. This is especially true in regard to shared research data that are bi-directionally linked to published outputs in articles, data repositories, and datasets. The explosion in the amount of data produced, combined with advances in data science, present enormous opportunities for big science. Heavily data-intensive, computational, and collaborative research forms the basis of open science across diverse disciplines, countries, and technologies. In this context, the use of industrial-level equipment is increasingly prevalent as part of the effort to introduce more sophisticated analyses into large-scale research and to maintain transparency and public trust in science by validating original research findings.

In a manner consistent with the open science movement, open access to research data are mandated by the major funding agencies (e.g., the National Science Foundation and National Institutes of Health), high profile journals (e.g., Nature) and data journals (e.g., the PLoS family of publications or Scientific Data). While the possibility of data scooping, theft (Borgman, 2007), or manipulation remains, research data are increasingly shared and made available to the public for reuse. In order to be in compliance with data sharing requirements, researchers are required to submit data such as computer codes and datasets as supplementary information. There are, however, barriers to open science in scholarly communication owing to the incomplete development of a culture of sharing and reuse when it comes to publishing and repositories. From a technical perspective, the federation of emerging data infrastructures for open science, such as common interfaces and data standards, with continuous maintenance and interoperability alignment and best practices for data are insufficient. From the perspective of data sharers, on the other hand, data scooping and planarization (Borgman, 2007) or loss of publication opportunities

(Reidpath & Allotey, 2001; Stanley & Stanley, 1988) may be real concerns. Researchers' individual perceptions that current rewards systems do not guarantee credit that translates into tenure, successful grant applications, and promotions may also be a source of reluctance to share data. For potential data reusers, collecting data themselves may prove more attractive than the reuse of shared data owing to the time and effort (Kim & Stanton, 2015) that may be necessary to subject others' published data to further analysis and to adjust preexisting frameworks.

Previous studies have not developed an integrated view of the various factors that influence data sharing and reuse, which can be categorized as institutional, individual, and those relating to information technology (IT) resources (Kim & Stanton, 2015). Kim and Stanton found that both institutional pressures and individual motivations play significant roles in data sharing behaviors across diverse scientific disciplines. Thus, the data sharing behaviors of STEM researchers can be facilitated by attitudinal, normative, and resource-related considerations. Factors associated with shared gene expression microarray data, for example, which relates to one of the STEM fields, pertain to authorship, publication, funding, and institution and domain environments (Piwowar, 2010). In the social sciences, factors influencing data reuse include the processing of trust judgments with various types and levels of trust interaction (Yoon, 2015) and the perceptions and practice of data reuse differ between novice scientists and experts (Faniel, Kriesberg, & Yakel, 2012).

The sharing of detailed descriptions of research data is associated with a 69% increase in citation rates (Piwowar, Day, & Fridsma, 2007). Data citation serves, among other things, to identify, authenticate, locate, access, and interpret published data as well as to give credit and to establish provenance (CODATA-ICSTI Task Group on Data Citation Practices, 2013). Common practices in research data citation have not yet, however, been broadly implemented that would accurately apportion credit, incentives, recognition, and rewards by means of bibliographic references to published research data; published research data are, as noted, regarded as part of the supplementary materials (CODATA-ICSTI Task Group on Data Citation Practices, 2013). In the open science movement, therefore, researchers may wish to increase the impact of their scholarship in terms of recognition and rewards that may accrue after their data have been shared (i.e., through data publication) and reused. Researchers receive more credit when researchers share their research data (Piwowar & Vision, 2013) and researchers are more inclined to share their research data if researchers receive more credit (Borgman, 2012). It is not, however, easy to measure the impact of data citation automatically owing to the lack of standards or guidelines for article citation that are universally accepted among publishers, journal editors, and funding agencies. The absence of uniquely identifiable research resources represents another limitation. Although principles of data citation have been articulated (Data Citation Synthesis Working Group, 2014), researchers remain hesitant, owing to the lack of clear standards, guidelines, or mechanisms in the peer review process (for both regular journals and data journals), to share their data with the public. Further, a recent study has reported the persistence of data citation of the self-cited variety (Park & Wolfram, 2017; Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2016), a situation that may hinder the potential for the future reuse of shared research data. It is also the case that the methodology for research data citation is still in its infancy; previous work has consisted of exploratory studies without any guiding methodological or theoretical framework or any proposals regarding the varying degree to which different factors impact data citation. The goal of this study is, accordingly, to identify and evaluate a reliable way of measuring the scholarly impact of the citation of open research data and a methodological framework for approaching questions relating to data sharing and reuse.

## 1.2. Significance of the Research

This dissertation makes several contributions to the existing nascent methodological and practical framework for research data citation. To begin with, the quantitative methods used here are intended to help explain the phenomenon of research data citation across diverse disciplines in broadly applicable terms. More specifically, the aim is to expand the understanding of data citation in the context of open science through exploratory research. Although previous work has assessed the impact of data reuse and sharing in the social sciences—based on the Inter-University Consortium for Political and Social Research (ICPSR) repository (Fear, 2013)—and in biomedical research (Piwowar, 2010), less attention has been directed to the actual extent and impact of data citation. Moreover, disagreement persists as to the meaning of data citation across scientific disciplines in scholarly communication, possibly owing to the complexity of the concept. For this reason, the exploration of the role of data citation in scholarly communication, especially in the context of the STEM fields, offered here is another potentially valuable contribution to the existing literature.

Similarly, valuable, from a methodological perspective, is the elaboration here of a framework for the study of data citation (i.e., in the context of data sharing and data reuse). The absence of such a framework is due to the fact that data citation, especially in data journals, is a relatively new phenomenon. Methodologies developed to study similar phenomena may prove applicable or may at least point the way to avenues for further research.

In more practical terms, findings presented here can provide insight into the impact of data citation by scientists, especially as it relates to the field of scientometrics in scholarly communication. A better grasp of the factors that impact data citation can in turn enhance the understanding of factors that determine the efficacy of scholarly communication, which are of concern for individual scientists and scientific institutions alike. At the same time, this research may prove useful in the development of guidelines, standards, and recommendations for improving current citation activities in the data management life cycle as well as of policies governing data citation for journal publishers, institutions of higher education, and funding agencies.

## 1.3. Research Questions and Purpose

In light of the impact of data citation, and in the context of the open science movement, the main purpose of this dissertation is to improve the manner and the extent to which the sharing and reuse of research data affect their citation in the STEM fields. Multiple disciplines have been selected because, in the era of big data and open science, in which large-scale research across diverse disciplines using industry-level equipment is commonplace, the impact of scientific data citation in general cannot be studied without considering specific disciplinary factors. Moreover, as discussed, the impact of data citation across disciplines, as revealed by scholarly databases, data repositories, and data journals, remains relatively unexplored from the perspectives of data sharing and reuse. Accordingly, the specific research questions addressed here are:

- RQ1: How prevalent is data sharing in different disciplines as measured by formal data citation in STEM fields?
- RQ2: What types of STEM research data are formally cited most often?
- RQ3: How do author self-citation/recitation practices differ across STEM disciplines?)?
- RQ4. How do data reuse practices differ across STEM disciplines?
- RQ5: To what extent do STEM disciplines support interdisciplinary data citation?

The first research question is intended to identify and map factors at various levels that influence the impact of the sharing and reuse of research data in the STEM fields generally. The second research question evaluates the impact of each factor identified in answering the first question on data citation, again in general. The first and second research questions are interconnected and expected to provide an integrated and refined view of the significance of data citation across STEM disciplines. The third research question seeks to identify factors associated with author self-citation or recitation. It is important to examine these phenomena as well as disciplinary factors (i.e., across disciplines) because they are fairly prevalent in research data citation (Park & Wolfram, 2017), while each discipline displays distinctive citation behavior (Helbig, Hausstein, & Toepfer, 2015; Torres-Salinas, Jiménez-Contreras, & Robinson-García, 2014). The fourth research question concerns data reuse practices and the fifth interdisciplinary data citation, again in the STEM fields.

The frequency of data sharing also varies within scientific communities (Tenopir et al., 2011). Thus, regarding self-citation, some authors tend to use the same shared research data repeatedly (Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2016), and author self-citation is likewise prevalent in research data citation in genetics and heredity (Park & Wolfram, 2017).

## 1.4. Scope

The analysis presented here does not extend to the social sciences and humanities. STEM fields have been early adopters of open science initiatives in comparison to social sciences and humanities and have more broadly adopted data sharing (Park & Wolfram, 2017). Further, this study does not include altmetrics derived from Google Scholar or such social network platforms as YouTube, Twitter, Facebook, or Google+ because “research data are either rarely published or not findable on social media platforms" (Peters, Kraker, Lex, Gumpenberger, & Gorraiz, 2016, p. 741) and because altmetrics scores for research data are very low (Peters, Kraker, Lex, Gumpenberger & Gorraiz, 2015). Altmetrics is a non-traditional and fairly new (first appearing in 2010) form of informetrics. Thus, the data are limited to records indexed in Clavariate Analytics' Web of Science and, in particular, the Data Citation Index (DCI).

## 1.5. Definition of Terms

*Data citation.* Data citation is the key practice that provides a reference to research data for its recognition as primary research results. Data citation broadly speaking involves credit, attribution, and discovery of data (Borgman, 2016). A reference to other publications such as journal articles or books to author's own primary data can also be regarded as a data citation.

*Data publishing.* Data publishing is the release of data in published formats for public use or reuse. The basic classes of data publication are journal-driven archival data, appendix data, standalone data publications, publication by proxy, and overlay publication (Lawrence, Jones, Mattews, Pepler, & Callaghan, 2011).

*Data sharing.* Data sharing refers to the "release of research data for use by others" (Borgman, 2012, p. 3) or the release of the raw/pre-processed or primary research data by researchers or institutions, whether voluntarily or in accordance with institutional norms (Curty , 2015). Data sharing is affected by the predilections of individuals and institutions (Kim & Stanton, 2015), or "a voluntary provision of information from one individual or institution to another for purposes of legitimate scientific research" (Borouch, 1985, p. 89), through central or local data repositories or personal communication methods (e.g., exchanges of data among acquaintances).

*Data reuse.* Data reuse is the use of existing data by scientists to replicate or reproduce outcomes of a previous study by combining with it other existing or newly collected data (King, 1995) obtained from repositories or through personal communication channels (e.g., acquaintances). Data reuse includes any secondary deployment of original or existing research data in order to study new problems; it generally represents different dimensions of, and cases described as, the secondary analysis of existing data (Curty, 2015).

*Formal data citation.* Formal data citation refers to instances in which data sharing and reuse are cited or described in the references section in addition to the main text of a paper in such a manner that the sharers of the data may receive due scholarly credit.

*Informal data citation.* Informal data citation refers to the sharing and reuse of data in contexts other than the formal references section and in such a manner that they cannot be readily indexed by a citation indexing service. Such citations may be located in the main text or the acknowledgments section of a paper and positioned such that, again, the sharers of the data are not likely to receive formal scholarly credit.

*Research data.* Research data include any form of data obtained by researchers that is accepted or retained in scholarly communication in order to produce original research outcomes or to validate research findings. These data include such information as research techniques and materials (Blumenthal et al., 2006); their types include raw or analyzed, observational, experimental, simulation, derived or compiled, and reference or canonical.

## 1.6. Dissertation Structure

This dissertation consists of six chapters. Chapter 1 explains the open science movement in scholarly communication from the discussion on data sharing and reuse on data citation. Chapter 2 surveys relevant findings and methods from previous studies and made clear the lack of comprehensive research on data citation, a gap that this research aims to fill. Chapter 3 describes and justifies the methodological approaches employed in this dissertation. Because this was relatively new ground, exploratory mixed methods were used in order to understand the factors that influence the impact of data citation on scientists in terms of data sharing and reuse. A set of factors found to affect or be affected by scientists' decisions regarding data citation that were based on its sharing and reuse were elaborated. Chapter 4 reports the research findings that emerged from the data analysis as well as the pilot study to identify additional indicating terms for data citation. Chapter 5 discusses important points that emerged from the research results and explains the limitations and implications of this study. Chapter 6 concludes the dissertation with a summary of the research findings and directions for future studies.

# Chapter 2 LITERATURE REVIEW

## 2.1. Introduction

This chapter reviewed research on relevant aspects of scientific communication and informetrics, and in particular citation analysis in scholarly communication and the literature on data citation, sharing and reuse by researchers. Scientific communication represents a subset of the larger field of scholarly communication. Traditionally, scientists have shared scholarly knowledge with each other through two basic channels, informal and formal. Acquaintances share new knowledge through such informal channels as email messages, conferences, and personal letters. Formal channels, on the other hand, include invisible colleges and formal scholarly publications, such as journal articles, monographs, and conference proceedings. An invisible college is “an elite of mutually interacting and productive scientists within a research area” (Crane, 1972, p. 34) that may not involve a permanent record. In order to examine the impact of data citation, this dissertation reviewed informetrics in scholarly communication and the open science movement as well as data citation/reuse. These findings were then synthesized, and the chapter concluded with a discussion of the limitations of previous research in relation to the research problems.

## 2.2. Metric Studies of Scientific Communication

### 2.2.1. Scientometrics

Informetrics is the quantitative investigation of forms of information production and their usage in recorded discourse (Tague-Sutchliffe, 1992), or “the structural relationships within the literature itself” (Wilson, 1999, p. 109). As a sub-domain of information science, it “covers and replaces the field of bibliometrics, including citation analysis, and includes some recent subfields such as Webometrics" (Wilson, 1999, p. 115). The production and use of knowledge are studied from scholarly and professional perspectives, both quantitatively and qualitatively, in order to examine the process of knowledge creation, dissemination, and implementation. Quantitative artifact approaches include informetrics and forms of social network representation. Qualitative and interpersonal approaches, such as the peer-review process, may incorporate social network analysis. Among the areas of study in informetrics are author productivity, journal productivity, citation and co-citation analysis, recorded language, the growth and obsolescence of literature, and the use of resources.

Allied "metric" areas of informetrics include bibliometrics, scientometrics, webometrics, cybermetrics, and altmetrics. Egghe (2005, p. 1311) defines informetrics as "the broad term comprising all the metrics studies related to information science, including bibliometrics (bibliographies, libraries . . .), scientometrics (science policy, citation analysis, research evaluation . . .), and webometrics (metrics of the web, the Internet or other social network such as citation or collaboration network)." Bibliometrics (Pritchard, 1969), the quantitative study of recorded discourse, represents the formalization of statistical bibliography. Webometrics qualifies as a subset of informetrics because, in the study of web phenomena (Björneborn & Ingwersen, 2004), hyperlinks are treated as citations; thus, link analysis is treated as citation analysis and co-link analysis is treated as co-citation analysis in the web environment. Webometrics and cybermetrics are often used synonymously, though Björneborn and Ingwersen consider the former to be a subfield of the latter. Altmetrics (alternate or alternative citation metrics) involves the analysis of data from social media, such as blogs, microblogs (e.g., Twitter), social bookmarking data, and other alternative electronic sources, in order to assess impact; it provides new ways to track author influence on the social and scholarly web (Priem & Costello, 2010) through such platforms as Mendeley or CiteULike.

With the advent of the former Institute for Scientific Information (ISI) citation indexes, the analysis of reasonably sized literatures without laborious data collection became possible (Wilson, 1999). Thus, rather than relying only on surrogates or bibliographic representation, further statistical analysis could now be performed on the articles themselves using information in digital form and access to the full text. Academic databases that provide citation index services include the Clarivate Analytics Web of Science (WoS; https://webofknowledge.com), Elsevier's SciVerse Scopus (https://www.scopus.com), and Google Scholar (http://scholar.google.com), the Korea Citation Index (KCI) of the National Research Foundation of Korea (NRF) (https://www.kci.go.kr/) and the Chinese Science Citation Database (CSCD) of the Chinese Academy of Sciences. These citation databases are key sources of data for citation analysis that help in the understanding of scholarly communication, the intellectual structure of disciplines, and the impact of research. Traditionally, informetric analyses have been used for the development of scientific indicators, library collection management, the development of science policy, and the design and evaluation of information systems. The areas of greatest relevance to the current research are scientometrics and citation analysis.

Scientometrics refers to "the quantitative study of science and technology" (Wilson, 1999, p. 110) and involves the empirical analysis and measurement of text or documents in the fields of science and technology in order to examine the patterns, structures, and behaviors of science and technology. Both quantitative and qualitative methods are necessary for scientometric analysis, with the emphasis on the former.

### 2.2.2. Citation Analysis

Citation analysis represents a core area of investigation in informetrics and scholarly communication. It deals with the relationships between a part or the whole of a cited work and a part or the whole of a citing work, including articles, authors, journals, or groups. Research findings from citation analysis can help to increase the understanding of scholarly communication and disciplinary relationships. Although the idea of citations goes back further, modern citation analysis traces back to the pioneering work of Eugene Garfield (1955), who initiated the development of the Science Citation Index (SCI). The concept of citation analysis formed the basis of informetrics, bibliometrics, scientometrics, and webometrics. Citation analysis is usually directional. Börner, Chen, & Boyak  (2003) have added definitions as follows:

> A "citation" is the referencing of a document by a more recently published document. The document making the citation is the "citing" document, and the one receiving the citation is the "cited" document. Citations may be counted and used as a threshold (e.g., only keep the documents that have been cited more than five times) in a mapping exercises. Other terms used to describe citing and cited numbers are "in-degree," or the number of times cited, and "out-degree,", or the number of items in a document's reference list.

Smith (1981) discussed the use of citation analysis to describe patterns of citation, evaluate influences and productivities, and facilitate document search and retrieval. Using citation, credit for multiple-authored works (e.g., equal count, first author count, fractional count, and proportional count) and self-citations can be addressed. Measurements of citation include citation count, adjusted citation count, citations per publication, and adjusted citations per publication. Citation data are often used to analyze the obsolescence of scholarly literature. The assumptions underlying citation analysis are that (a) the citation of a document implies its use by the citing author; (b) citation implies that a document has such merits as quality, significance, and impact; (c) citations are made to the best possible works; (d) a cited work is related in content to the citing work; and (e) all citations are equal. Citations are essential in scientific communication.

Self-citation is the cited references of an author name that matches the name of the author of a citing article. Examples of self-citation are direct self-citation, author self-citation, and journal self-citation. Direct self-citation happens when the author cites his/her previous works subsequently in scholarly works. Author self-citation (i.e., direct self-citations for the co-author(s)) happens if one or more co-author(s) of researcher *A* publish another work without researcher *A* and that other work (i.e., paper) cites their work (Glänzel & Thijs, 2004). Although self-citations do not only explain higher impact of collaborative papers (Van Raan, 1998), in big science, the possibilities of author self-citation arise due to the large numbers of co-authors in each publication (e.g., journal) where large densely connected collaborative research teams across multiple disciplines. Journal self-citations represent how often a journal (i.e., a work) is cited by its own publications (i.e., journal) (Leydesdorff, 2008). Journal self-citations can manipulate the journal impact factor (Krauss, 2007) from the Journal Citation Reports (JCR). The JCR, provided in conjunction with the WoS since 1975, is a source of information about the impact and influences of scholarly works. The idea of a journal impact factor is discussed in (Section 2.2.8).

Citations can provide appropriate acknowledgement, rewards, and justification for researchers' findings (Latour, 1987). Researchers in scholarly communication obtain recognition by publishing research and being cited in other research, because, as Borgman (2007) mentions, the rewards system approaches the research publication as a scholarly communication practice.

### 2.2.3. Citation Counts

Merton (1968), in observing that scientists use citations in order to give due credit to each other, described what is often referred to as the normative theory of citation. The contrasting constructivist approach proposes that citations serve other purposes, including advancing scientists' interests and defending their claims, persuading other people, and establishing a position in a scientific discussion (Brooks, 1986; Gilbert, 2015; Moed, Glänzel, & Schmoch, 2004). Observing this debate, Lawani and Bayer (1983) have suggested that, despite ambiguities in citation practices, considerable evidence has accumulated to suggest that citations provide an objective measure of what may be termed scientists' "productivity," "significance," "quality," "utility," "influence," "impact," or "effectiveness." In any case, the debate continues despite attempts to get beyond it (Cronin, 1984, p. 103; Cronin, 2014, pp. 3-21; Moed, 2005, p. 346; Wouters, 1999).

For the purposes of this study, citation count can be expressed as

$$\mathrm{C} = \{C_1, C_2, C_3 \ldots, C_p\}$$

$$C_1 \geq C_2 \geq C_3 \ldots \geq C_P \wedge \sum_{i=1}^{P} C_i = C_T \wedge C_1 = max_i(C_i)$$

Citation count, or the total number of citations, is $C_T$; P is the number of papers of an author, C is the set of the citations received by the P ordered papers. Citation sequence or citation profile $\{C_1, C_2, C_3 \ldots, C_p\}$ can be used to characterize an author. $C_1$ is the number of citations received by the most cited research output.

The most common measure, for both the productivity and the impact of the performance of an author (journal, group of authors, institution or country), may thus be the total number of citations of all papers published.

### 2.2.4. Direct Citation

Direct citation, sometimes referred to as cross citation, can serve as a measure of the relatedness between cited and citing works. It is defined as “the citing of an earlier document by a new document” (Small, 1973, p. 265), meaning that author *A* directly cites author *B*. Direct citation has not been studied actively so far, though it is now beginning to receive some attention (Boyack & Klavans, 2010) through the comparison of direct citation with other forms of citation-based measures such as co-citations and bibliographic coupling. Boyack and Klavans have revealed that direct citations are less accurate than co-citations and bibliographic coupling slightly outperforms co-citation analysis. Direct citations are not actively used for visualizations, but rather indicate more direct publication relations.

The uniform direct citation of data curated in persistent data repositories has received attention because “a foundational element of reproducibility and reusability is the open and persistently available presentation of research data” (Starr et al., 2015). Assigning a permanent and persistent data identifier, such as a Digital Object Identifier (DOI), at the time of data publication may be important for direct citation and ease of accessibility in data citation because a DOI is machine-readable and therefore provides access to cited data and its associated metadata (Borgman, 2016). However, current practices are such that data citation possesses a low percentage of persistent identifiers (e.g., DOI, Open Researcher and Contributor IDentifier; ORCID) compared with regular citation. Access to data repositories (e.g., open access data repositories), whether unrestricted, limited, or restricted, needs to be studied in the context of the sharing and potential reuse of data.

With regard to mapping scholarly literature, direct citation relations among publications have less often served as a tool for visualization than co-citation, possibly because direct citation often leads to networks with only a small number of edges (van Eck & Waltman, 2014). Direct citation played an important role in Eugene Garfield's work on algorithmic historiography, and the publication citation networks of direct citations can be mapped using his HistCite tool (http://interest.science.thomsonreuters.com/forms/HistCite/) which takes WoS output file formats as input that is visualized as a historiograph. CitNetExplore (http://www.citnetexplorer.nl) is a similar tool that can be used to map more extensive analyses, while CiteSpace (http://cluster.cis.drexel.edu/~cchen/citespace) is a tool for visualizing patterns and trends in scientific literature. As Cobo, Lopez-Herrera, Herrera-Viedma and Herrerea (2011) observe, a relationship between units can be established using direct linkages. Examples include a document-document, author-author, or journal-journal citation network.

### 2.2.5. Co-citation and Literature Mapping

Co-citation, usually in the form of bidirectional citation analysis, is a measure of the common occurrence of two entities of interest (e.g., publications, authors, or journals) in the reference list of a third document. It determines the semantic similarity among documents regarding citation relationships based on the frequency of co-citations and can thus be described as "an interpretation of the significance of strong co-citation links [that] must rely both on the notion of subject similarity and on the association or co-occurrence of ideas" (Small, 1973). White and McCain (1997, p. 103) have defined measures of co-citation as follows:

> Co-citation occurs when any two works appear in the references of a third work. The authors of the two co-cited works are co-cited authors. If the co-cited works appeared in two different journals, the latter are co-cited journals. Co-words are words that appear together in some piece of natural language, such as a title or abstract.

Co-citation is a generally accepted way to obtain relational information about documents within a domain (Moya-Anegon et al., 2004). Co-occurrences of citations (authors and papers) reveal relationships in “bibliographic coupling” and co-citation analysis (author, journal, and publication). In order to study co-citation, cluster analysis, multi-dimensional scaling (MDS), factor analysis and social network analysis may be applied. Among the limitations of co-citation are (a) possible omission of authors other than the first for a given work, (b) author ambiguity, and (c) sources with large numbers of references that may co-cite a great many documents. Co-citation is thus the opposite of bibliographic coupling. Citer-based analysis represents an alternative way to transcend some of the limitations of co-citation analysis (see below for further discussion).

Scientific domains have been studied using co-citation analysis, which was introduced by Small (1973) and Marshakova (1973) as a means to measure relatedness between pairs of documents or authors cited together. Co-citation of thematic or schematic representations of classifications (classes and categories) has been studied, and the mapping of large scientific domains has been identified as a significant method in this regard (Moya-Anegon et al., 2004). Co-citation networks in information science have also been studied using visualizations of the distance between two nodes, where relatedness is inversely proportional to the distance between them (White & McCain, 1998). Author co-citation analysis (ACA) (White & Griffith, 1980) has been applied to information retrieval (IR) and various other domains.

Document co-citation is used as a variable in order to build domain maps by analyzing citations of scientific production (Small, 1973). Domain analysis from the perspective of society, rather than that of pure abstract research, was introduced as a new method (Hjørland & Albrechtsen, 1995). A domain analytic technique was applied by White and McCain (1998) to visualize a discipline using co-citation in the field of information science. Co-citation is discussed as a method or tool for representing schematically domains that provide different viewpoints with respect to existing relationships among variables (e.g., authors, documents, journals, and words) (Moya-Anegon et al., 2004). Domain analysis is "the activity, or the methodology, by which the conceptual content and natural or heuristic ordering can be discovered and mapped in discrete knowledge domains" (Smiraglia, 2014, p. 85). Smiraglia (2002) used meta-analysis as a tool for knowledge organization. Smiraglia (2012) also used domain analysis as a tool to extract ontology for knowledge organization systems and to provide interoperability across diverse domains.

Boyak, Klavans and Börner (2005) note that Pearson correlation analysis has been used to analyze co-citation counts within articles about MDS in order to study a single discipline. The correlation is determined based on mapping citations within published journals. Moya-Anegon, et al. (2004) have employed co-cited ISI category assignments to create category maps as an alternative to using journals to map the structure of science. Boyak, Klavans and Börner have charted the whole of science by mapping over 7,000 journals from both the SCI and the Social Science Citation Index (SSCI) based on the notion that journal sets are associated with disciplines; similarity measures were based on journal inter-citation and co-citation frequencies. These researchers did not use data from the JCR because, while it contains inter-citation frequencies, "co-citation frequencies based on paper-level co-occurrences of references cannot be derived from anything but the original reference lists. (p.355)"

### 2.2.6. Bibliographic Coupling

Bibliographic coupling, proposed over half a century ago by Kessler (1963), occurs when two documents reference a common third document in their bibliographies, thus suggesting that they deal with similar subject matter. The degree of bibliographic coupling between citing documents is determined based on the percentage of total citations in common. It is mostly studied in the context of coupling analysis in informetrics and can be viewed as the inverse of the co-citation link because “bibliographic coupling is a technique for clustering (citing) documents according to their number of shared references. Co-citation analysis, on the other hand, is a technique for clustering (cited) documents according to their number of co-occurrences in subsequent documents’ reference lists” (Wilson, 1999, p. 148).

Author bibliographic coupling analysis (ABCA) extends bibliographic coupling to an author-aggregated approach on the document level (Zhao & Strotmann, 2008). Other types of author coupling, in addition to author bibliographic coupling, include author journal coupling, author keyword coupling, and author title-word coupling. In order to develop a technique for scientific mapping, bibliographic coupling can be used in combination with cluster analysis in cases in which bibliographic coupling and document co-citations are compared for the purpose of literature mapping (Jarneving, 2005; Jarneving, 2007).

### 2.2.7. Scholarly Impact Assessment

The impact factor (IF), so designated by Garfield and Sher (1963), is a measure that evaluates journals in combination with other measures and evaluations. The impact factor has been used to assess scholarly contributions, especially in the context of the WoS citation indexes (Wilson, 1999). Producers of scholarly contributions include individuals, departments, institutions, disciplines, and countries. Scholarly impact assessment is vital from the academic’s perspective, for, as assessed based on publications, presentations, and grants, it plays a significant role in the advancement and maintenance of careers. Traditional individual and institutional assessment measures include publication and citation counts and grant-seeking success.

Challenges to assessing scholarly impact assessment may include counting publications, the use of citations as a measurement unit, dissemination outlets, the Matthew effect (Merton, 1968; 1988), and the Podunk effect (Gaston, 1978). Counting publications may prove problematic owing to increasing levels of collaboration in scientific communications in the era of multi-authored works, big data, and open science. Another concern is exactly which publications are to be counted. The use of citations as a measurement unit is complicated when the number of publications from a single research project proliferates owing to the circulation of multiple versions of evolving research. Thus, for instance, a single research project may be represented variously by a work in progress poster, conference proceedings, and expanded articles in a refereed journal. When it comes to dissemination outlets, considerations include peer-reviewed versus non-peer-reviewed works and the treatment of OA journals, institutional repositories, project websites, and academic blogs. Lastly, the Matthew (Merton, 1968; 1988) and Podunk effects (Gaston, 1978) describe how authors may receive more or less credit than they deserve on account of their reputations or geographical locations, respectively; both effects are difficult to assess.

Impact factors have long invited debate, and they have been criticized on the grounds that "evaluations cannot be made with numbers in isolation if the basis (or unit) of comparison is uncertain" (Wilson, 1999, p. 131). Specific criticisms include the need to aggregate a set of documents (Egghe & Rousseau, 1996; Seglen, 1992; 1994), variation in citedness within a journal (Harter & Nisonger, 1997; Moed, Van Leeuwen, & Reedijk, 1999; Schwartz, 1997), appreciable differences in citedness among different disciplines for the same document type (Schubert & Braun, 1993; Schwartz, 1997), and the comparability of different units regarding the number of citations received (Moed, Van Leeuwen, & Reedijk, 1999).

Taking these criticisms in turn, with regard to the aggregation of a set of documents, the difficulty involves the highly skewed distribution of citations across articles, journals, and databases. The issue of variation in citedness within a journal arises owing to the inclusion of various document types with differing capacities to attract citations (Wilson, 1999). As Moed, Van, Leeuwen and Reedijk (1999) have noted, impact factors differ in biomedical areas depending on the document type. Thus, for example, articles, reviews, and notes were found to have higher impact values than editorials and letters. These researchers noted that, when the latter types of documents are included, journal impact factors may be 10 to 40% lower. Schwartz's (1997) analysis of different levels of citedness in different document types in the WoS databases revealed that 47% were uncited in the physical sciences, while the figure was 22% when conference abstracts, editorials, reviews and letters were excluded. Schwartz's findings are also relevant regarding appreciable differences in citedness among disciplines for the same document type (Wilson, 1999).

### 2.2.8. Journal Impact Factor

The journal impact factor (JIF), or simply impact factor (IF), has become the most popular and discussed approach to assessing the visibility and diffusion of journals in the period since it was first used in 1963 in the SCI (Garfield & Sher, 1963), during which it was reconstructed by Garfield (2006) and Archambault and Lariviere (2009). Gross and Gross (1927) initiated the use of references to assess scientific journals, while Eugene Garfield (1955) suggested that journal impact can be assessed based on counting references to journals.

Clarivate Analytics' JCR publishes the JIFs for thousands of journals annually, which is based on the journal citation itself. The JIF is based on publications and citations and can be used to compare the importance of sources based on citations received. The JIF is also called the two-year impact factor. Mathematically, it can be expressed as

$$IF_{j.y}^{(2)} = \frac{\sum_{i=1}^{2} C_j \, (y,y-i)}{\sum_{i=1}^{2} P_j \, (y-1)}$$

Where $IF^2$ is the two-year impact factor, $C_j(y, y-1)$ is the number of citations received in the year $y$ by articles published in journal $j$ in the year (y-1), and $P_j(y-1)$ is the number of articles published in journal $j$ in the year ($y$-1).

Tools for assessing journals include Journal Citation Reports, Eigenfactor, and the SCImago Journal Rank. Journal Citation Reports are based on the Clarivate Analytics WoS; Eigenfactor makes comparative measurements based on article influence and cost effectiveness; and the SCImago Journal Rank is based on Elsevier's Scopus data.

### 2.2.9. Co-word Analysis

Co-word analysis (Callon, Courtial, & Laville, 1991), also known as semantic mapping, uses language modeling and text mining approaches and the most important words or keywords of documents in order to study the conceptual structure of a domain. Co-word analysis accounts for (a) words/terms that occur with one another as a way to identify synonyms, (b) the relatedness that can be directly interpreted based on document contents, and (c) the frequency distribution of co-occurring words that, when tallied, follow a pattern similar to such other informetric regularities as long-tail distribution. The feasibility of co-word analysis as a method has also been studied (Ding, Chowdhury, & Foo, 2001). Co-word and co-citation analysis are similar in that both are used to determine the strength of relationships among textual containers and to identify similarities among the techniques used, such as cluster analyses and MDS methods.

The co-occurrences of keywords in articles have been used as an indication of associated strengths in order to map keyword relatedness. The benefit of co-word analysis is that a direct interpretation of relatedness is available based on document content. Weaknesses of co-word analysis include the fact that the meanings of words change over time and depending on context both within and among texts (Leydesdorff, 1997) as well as the indexer effect (Law & Whittaker, 1992), which results in delayed changes, the creation of bias, and the introduction of subjectivity into the index terms. In the words of Bhattacharya and Basu (1998), co-word structure can stand for research activities within scientific research and this approach has accordingly been applied to mapping scholarly literatures within a given research area at the micro-level.

### 2.2.10. Citer-based Analysis

The definition of self-citation has been "extended to include citations originating from publications authored by one of the coauthors of the cited publication of interest, or coauthor self-citations," in the words of Ajiferuke, Lu, and Wolfram (2010, p. 3). These researchers have discussed how citation counts largely include recitations (i.e., repeated citations by an author of the same work over time), for which reason they suggest using the citer (i.e., the origin of citation) as the unit of measure. Author-level recitation can be measured using the "analyze results" feature of the WoS, which provides lists of reciting authors and their frequencies. These researchers also mention as citer-based measures: citer count, citers per publication, and the ch-index. Citer count refers to the number of authors who have cited a publication by given author; citers per publication refers to the number of citers by the number of publications by an author; and the ch-index corresponds to *x* publications with at least *y* citers. Limitations of citer-based analysis include the fact that (a) most citer data are not easily extracted through current end-user interfaces and will not be so until and unless more raw data or more sophisticated queries become available, and that (b) this form of analysis is still based on the citation for measurements.

The consideration of hyperauthorship (Cronin, 1984) is necessary in studying self-citation because with hyper-authored works, the likelihood of self-citation increases due to there being more co-authors who are in a position to self-cite. Hyperauthorship is the practice of publishing papers with large numbers of co-authors, potentially hundreds. In interdisciplinary research for big sciences, hyperauthorship is common in some areas such as the hard sciences. For instance, the total number of authors in a given publication in high-energy physics can occasionally exceed 100 authors (Tarnow, 2002). By using citer-based analysis, Park and Wolfram (2017) found that author self-citation or recitation is prevalent for research data citation in genetics and heredity, meaning a small number of highly cited authors may be increasingly influential in data citation and an increase in citations does not necessarily indicate unique and new citers. The rates of self-citation were very low (1.2%) for traditional citation-based self-citation (i.e., bibliographic self-citation) but was higher (8%) in data citation.

## 2.3. Open Science

The open science movement works as the ground movement for data sharing and reuse by more concrete actions such as infrastructure and policies. Examples of necessary infrastructure include sustainable preservation and access to research data. Examples of policies include major funding agencies' or high impact journals' data sharing requirements. Borgman (2007) mentions that the combination of the open science initiatives and technological capabilities reconstructs scholarly communication in the digital age. Open science can be divided into three interdependent elements: open access, open data, open software (Peters & Roberts, 2012; Willinsky, 2005), open peer review and open notes. Transparency is important in scientific research because methods and results of a published study need to be accessible for detailed scrutiny.

There is both international and national support for the open access system. An example at the international level is the European Commission's Open Data for Europe, which states that open data are useful for funding agencies and patent services in an open access environment (European Commission, 2011). An example at the national level is the Royal Society in the United Kingdom, which promotes openness and transparency and infrastructures that meet standards of accessibility and intelligibility (Boulton et al., 2012).

### 2.3.1. Open Access

Open access is a communication channel in scholarly communication through which content can be accessed on their web site by the general public without financial or legal barriers for research purposes for any users to read, copy, download and use. It is this concept of free-of-charge access for the public that distinguishes OA journals from non-OA journals (i.e., traditional subscription-based journals). The difference extends to the financial infrastructure. Under a subscription-based infrastructure, authors and institutions are asked to pay the cost of the dissemination and use of scholarly knowledge.

### 2.3.2. Open Access Journals

OA journals are peer reviewed publications of scholarly communication that are freely available through the Internet and that generally allow authors to retain copyright. OA journals usually waive access fees, such as charging authors when a manuscript is accepted for publication; fees are usually waived or paid by author-sponsors rather than by the authors themselves. Regarding the cost of peer review and dissemination, OA journals have lower barriers to access compared with subscription-based journals (i.e., non-OA journals), the latter having no disciplinary repositories or data repositories and no peer-review process because of pre-prints and or post-prints. Authors hold copyright on these materials and their permission is required, whether for dissertations, course materials, or any other kinds of digital files. The Directory of Open Access Journals (DOAJ) indexes 12,134 OA journals from 123 countries as of September 2018 (Infrastructure Service for Open Access, 2018).

OA journals with shared data increase citation rate of articles. For instance, previous studies report that OA journals with their research data available have shown greater citation impact (Craig, Plume, McVeigh, Pingle, & Amin, 2007; Eysenbach, 2006; Norris, Oppenheim, & Rowland, 2008)

Advantages of OA journals include increased citation rates compared with traditional subscription-based journals (Harnad & Brody, 2004), opportunities to accelerate the review and publication process, and increased accessibility. Previous studies have compared the impact of OA and non-OA articles and found that the former have a considerably higher impact, at least in the context of citation counts in physics (Harnad & Brody, 2004). OA articles are more often cited than non-OA articles published in Proceedings of the National Academy of Sciences (PNAS), with the effect becoming more pronounced over time (Eysenbach, 2006). Other researchers have found the opposite, however, namely that the advantage of early access (i.e. early access effect) diminishes over time (Brody, Harnad, & Carr, 2006)..

Disadvantages of OA journals include the fact that they are not free of charge to the general public, the ultimate end-users of research output, and that most are held in rather low regard in scholarly communication. And while the cost of OA journals may be lower than that of non-OA journals, there are still significant expenses involved with the peer-review process and the production of a publication (Suber, 2002). Owing to the relatively low standing of OA journals, relatively little weight may be given to publications in OA journals with regard to a scholar's career advancement.

### 2.3.3. Open Peer Review

Open peer review (OPR), though it has yet to be widely adopted, is an emerging approach to peer review in scholarly communication in the context of the open science movement. Wang et al. (2016) noted that the process of OPR involves the evaluation of research by peer reviewers in order to identify flaws in research and to determine whether it meets established standards. OPR makes scientific discoveries open and transparent, meaning that the content of peer review is publicly available for scientific communication. Examples of OPR are Faculty of 1000 (F1000; www.f1000research.com/) and PeerJ (https://peerj.com/). F1000 is an example that adopted full OPR for open publication and open evaluation (OE) for life scientists and clinical researchers. PeerJ is an example that adopted an optional OPR, meaning a blind review process followed by optional publication of review history. With the analysis of one optional OPR Journal, PeerJ, Wang, You, Rath and Wolfram (2016) found that authors are still reluctant to make their reviews publicly available and for reviewers to identify themselves.

### 2.3.4. Open Data

The term “open data” is widely used in the scientific, governmental, and industry sectors. In this study, open data are data utilized in a scientific context to which scientists have access for reuse, including secondary analysis. To be considered open, data should be free of charge and freely available to the general public. For instance, Google announced *Google Dataset Search* to support and promote the sharing of open data across the Web by using a simple keyword search (Google, 2018). Data repositories and data centers thus represent core infrastructure when it comes to increasing access to research data. Open data in scholarly communication may include such research outputs as datasets of various sizes and formats, software codes, analysis code, and any technical environments needed to process the data.

Mauthner (2012) states that the 1950s represented the beginning of open data in the sciences with the early World Data Centers for geophysical sciences, followed by databases to archive Deoxyribonucleic Acid (DNA) sequences in the 1980s, the treatment of research data as public and sharable data in the Natural Sciences in the 1990s and more concrete regulations for open government data in the early 2010s.

Open data mandates from the open science movement that have been established by a number of countries, journals, and major funding agencies demand that research data be made available to the general public for use by other researchers. These mandates have accelerated the free exchange of research data in open science. Examples include the U.S. government policy on open data, the National Science Foundation (NSF) data management plan, and the data sharing policy of the National Institutes of Health (NIH). Thus the U.S. government policy (Executive Office of the President, 2013) has made open government data available to the general public to access and reuse, and the NIH’s data sharing policy has, since 2003, made final research data as widely and freely available as possible for research purposes while protecting confidential and proprietary data and safeguarding the privacy of participants (National Institutes of Health, 2003). Likewise, the NSF has since 2011 mandated the inclusion of a supplementary document containing a data management plan for the dissemination and sharing research results to the public (National Science Foundation, 2011). Other countries such as the United Kingdom require a data management plan for data sharing for the grantees of the Economic and Social Research Council (Economic and Social Research Council, 2015) and the European Commission has its data management plan by linking dataset, research publications, and author information (European Commission, 2016). Some journals, such as data journals (e.g., Data Science Journal), mandate that authors share their research data. Authors can choose whether their data will be made publicly available at the time of publication or instead after an embargo period. However, some scientists may not follow data sharing mandates because of the lack of enforcement mechanisms (Piwowar, 2010). Journal policy for research data has been announced in 2014 by *PLoS ONE* (Silva, 2014) and in 2017 by *Nature Publishing group* (2017), *Science* (2017) and *Elsevier* (2017).

The benefits of data sharing in an open science paradigm may include increasing data discoverability and accessibility, facilitating interdisciplinary research in scholarly communication, and providing greater transparency and openness in science. Defining such intellectual property issues as copyright, ownership, authorship, and responsibilities is an important part of formal data sharing when it comes to controlling ethical violations and scientific misconduct (Wallis & Borgman, 2011). Borgman (2012) has detailed beneficial aspects of data sharing in terms of four rationales: to reproduce or verify, to make the outcomes of publicly funded research open to the general public, to ask new research questions using existent data, and to advance the state of research. As open data increase the rate of bibliographic citation (Piwowar, Day, & Fridsman, 2007), open data also makes datasets citable and data citation possible. The detailed description of research data is associated with increased bibliographic citation rate (Piwowar, 2010) with citation benefits from open data (Piwowar & Vision, 2013). This indicates when the detailed description of open data is provided, the datasets are citable by making data citation possible.

## 2.4. Data Sharing, Reuse and Citation

### 2.4.1. Data Sharing

Major funding agencies now require a data sharing policy; the NIH since 2003 and the NSF since 2011. High profile journals such as *Nature Physics* require permission for authors to share their research data, whether public sharing at the time of publication or after embargoed period. In aligning with these requirements, researchers need to submit their research data in the form of datasets or software.

Data sharing can help more researchers receive rewards from researchers' shared data. Previous studies found that researchers withhold their research data rather than sharing in journals (Campbell & Bendavid, 2003; Cohen, 1995; Piwowar, 2011). Researchers tend not to share their data if low or no rewards are perceived for data sharers (Sterlling & Weinkam, 1990) although researchers' perceptions and rewards enhance data sharing behaviors (Kling & Spector, 2003). Researchers in scholarly communications perceive that current reward systems do not provide sufficient rewards or credits toward promotion, social recognition, successful grant applications and tenure (Kim, 2013). In STEM, researchers are reluctant for data sharing because of the concerns for lack of rewards and credits, data misuse or misinterpretation and too much effort with very few perceived returns (Kim & Stanton, 2015; Tenopir et al., 2011). In social science, researchers confront high ethical standards by social science communities (Israel & Hay, 2006) and data sharing and reuse in social sciences are often regarded as too complex due to the high probabilities of using qualitative data (Yoon, 2014). A previous study examined social scientists' data sharing behaviors (Kim & Adler, 2015). Although this study examined the pressures from funding agencies and journal publishers would influence social scientists' data sharing behaviors, no statistical evidences are found.

Two types of data sharing practices, formal data sharing and informal data sharing, were classified by Clubb and colleagues (1985) in the mid-1980s. Formal data sharing occurs in a structured way that involves intermediary channels that function as local or central repositories and dissemination services such as academic institutions. Informal data sharing occurs among the same area or discipline members usually in the form of copies of datasets or upon individual request, or more ad hoc ways. Regarding formal data sharing, Clubb and colleagues mentioned the advantages as broad data accessibility, which facilitates the interdisciplinary research because data are formally shared by repositories such as academic institutions. Regarding informal data sharing, advantages include (1) the high trust and low risk perception among involved individuals and (2) low immediate cost due to the absence of intermediaries.

In STEM fields, until recently, data sharing has been studied at the individual discipline level. However, in a modern science where collaborations across labs, department, colleges or even countries are commonplace, without considering disciplinary differences, data sharing in scientific disciplines in general cannot be studied. For that reason, Kim (2013) examined scientists' data sharing behaviors in multiple scientific disciplines, with the examination of institutional and individual influences.

In the social sciences, data sharing is not new. Social science data are context-based and involves direct or indirect interactions with human subject. Social scientists used others' shared data for the original study verification or for the reanalysis and producing new research in the 1970s and early 1980s (Feinberg, Martin, & Straf, 1985). Social scientists tend to more concerned about data misuse by others than STEM disciplines (Tenopir et al., 2011). For instance, Tenopir found that 23% of researchers (47 out of 204 surveys) agreed or somewhat agreed to easy access of their research data. In contrast, 49% of researchers in biology agreed and somewhat agreed to data sharing, which is almost two times higher in biology than the social sciences. However, in interdisciplinary domains, social scientists showed positive attitudes regarding interdisciplinary data sharing, such as anthropology combining with the earth and environment by using time-series remote sensing research data (White, 1991).

Qualitative research data sharing and archiving are increasing (Rasmussen, 2011). Qualitative data sharing is regarded as more complex than quantitative data sharing (Bishop, 2009) and often regarded as too complex for data sharing and reuse (Yoon, 2014). Direct and indirect interactions with human involvement can bring ethical concerns, especially for qualitative data. Ethical concerns such as sensitive personal information (e.g., protecting participants' identity before preserved in digital repositories) make qualitative data sharing complicated. Due to these reasons, there is persistent skepticism by today's researchers for qualitative data sharing and reuse (Mason, 2007; Mauthner & Parry, 2009; Slavnic, 2011; Yoon, 2014). In qualitative data sharing, challenges usually center on the methodology, due to the subjectivity in qualitative methodology (Bishop, 2005; Mauthner & Parry, 2009; Parry & Mauthner, 2004). The benefits of qualitative data sharing include reanalysis and reinterpretation.

### 2.4.2. Data Reuse

Different quantitative and qualitative methods have been used to study data reuse. These are summarized in Table 1. Researchers applied survey, statistical analysis or citation analysis for quantitative study. Researchers applied interview, content analysis, case study or ethnography for qualitative methods. Regarding qualitative study, Daniels (2014) applied comparative case study methods to exploratory study using semi-structured interviews and non-participant observation with purposive sampling. In this study, data were collected by using semi-structured interviews, nonparticipant observation, and research with historical records. Depending on the participants, the various data collection methods included: (1) interview, concurrent; (2) interview, retrospective; and (3) observation and interview, concurrent. Daniels employed for data analysis (1) iterative thematic coding of interview transcripts and (2) observation notes using the qualitative coding software NVivo.

Relatively recently, researchers used mixed methods for data reuse. One such example is Curty (2015), who has employed a mixed-method approach that combines a quantitative survey instrument and a qualitative interview instrument in order to identify factors that influence data reuse among social scientists. Likewise, Tenopir et al. (2015) used both quantitative survey methods with close-ended questions and qualitative ethnography methods in order to study changes and differences in practices and perceptions of data sharing and data reuse (1) among research scientists world-wide and (2) across geographic regions, age groups and subject disciplines. These researchers employed snowball and volunteer sampling methods in order to recruit participants. Park and Wolfram (2017) used both quantitative citation analysis and qualitative content analysis to examine the practices of data reuse and sharing on data citation in Genetics and Heredity. Park, You and Wolfram (2018) used both quantitative descriptive analysis and qualitative content analysis to examine the current practices of data reuse and sharing on data citation in biomedical fields.

As seen in Table 1, exploratory methods are actively used in data reuse (Curty, 2015; Daniels, 2014; Park & Wolfram, 2017; Park, You, & Wolfram, 2018). Thus, Curty has employed the exploratory sequential approach among her mixed methods by using interviews as a qualitative instrument and an online survey as a quantitative instrument. Her qualitative data collection and analysis process involves (1) a small-scale study; (2) interviews; (3) complementing cutting-edge academic literature; (4) exploring the research phenomenon; and (5) grounding preliminary findings in a research framework. In terms of quantitative data collection and analysis, Curty uses (1) a survey study with a larger group of social scientists and (2) testing of the research model and hypothesis.

Table 1 Methods used for the study of data reuse

| **type** | **quantitative methods** | **qualitative methods** | **source** |
|---|---|---|---|
| **data reuse** | survey | - | (Curty, Crowston, Specht, Grant, & Dalton, 2017; Joo, Kim, & Kim, 2017; Joo & Kim, 2017; Kim & Yoon, 2017) |
| | survey | interview | (Curty, 2015) |
| | - | case study, interviews | (Daniels, 2014) |

| | | | |
|---|---|---|---|
| | - | interview | (Curty, 2016; Faniel, Kriesberg, & Yakel, 2012; Faniel & Jacobsen, 2010; Rolland & Lee, 2013) |
| **data reuse /sharing** | survey | ethnography | (Tenopir et al., 2015) |
| | - | interview | (Dallmeier-Tiessen et al., 2014; Zimmerman, 2008) |
| | - | interview, ethnography | (Wallis, Rolando, & Borgman, 2013) |
| **data reuse/ curation** | - | interview | (Yoon, 2015; 2017) |
| **data reuse/ citation** | citation analysis | - | (He & Nahar, 2016) |
| **data reuse /sharing /citation** | statistical analysis | - | (Piwowar & Vision, 2013) |
| | citation analysis, citer-based analysis, descriptive analysis | content analysis | (Park & Wolfram, 2017) |
| | descriptive analysis | content analysis | (Park, You, & Wolfram, 2018) |

Data reuse across multiple scholarly communities has not yet been widely studied as a domain, though the social sciences have been actively approached from this perspective, using mainly interviews as the instrument (Curty, 2015; Daniels, 2014; Yoon, 2015). Examples include studies of social scientists' trust judgments regarding data reuse (Yoon, 2015), impact measurements of data reuse in social science (Fear, 2013), factors influencing research data reuse in social science (Curty, 2015), and data reuse in the context of museums (Daniels, 2014).

Data reuse across multiple communities in STEM fields is important because interdisciplinary research is necessary to address today's complex research problems. Thus, for example, as Jirotka and colleagues (2005) have discussed, a national database of mammogram images can be useful to epidemiologists exploring factors that contribute to breast cancer. However, disparities across disciplines create difficulties, particularly in regard to terminology (Pierce, 1999). Moreover, it is not clear that scientists (e.g., those in the STEM fields) are interested in reusing data collected by non-scientists (e.g., humanities).

Qualitative data reuse is important to consider. Hinds, Vogel, and Clarke-Steffen (1997) have identified both data-specific and general methodological challenges that must be overcome for the reuse of qualitative datasets. The latter includes "the degree to which the data generated by individual qualitative methods are amenable to a secondary analysis and the extent to which the research purpose of the secondary analysis can differ from that of the primary study without invalidating the effort and the findings" (Hinds, Vogel, & Clarke-Steffen, 1997, p. 411). Among the challenges specific to data sets are obtaining informed consent from participants in primary studies for data reuse and assessing the nature and quality of a qualitative dataset from original studies (Hinds, Vogel, & Clarke-Steffen, 1997). Especially for the qualitative data reuse among social scientists, trust judgment issues and validity of data are important for data reusers (Yoon, 2015).

### 2.4.3. Data Citation

The reward system in scholarly communication is traditionally based on the research impact in part. The research impact is based on the publications in peer-reviewed journals and the impact of those published journal articles. The establishment of formal data citation practices is needed to create new incentives as a parallel to current reward systems. Data citation is expected to create data stewardship and enhance data sharing as well as make research data more accessible and exploitable. Although data citation practices are not (yet) widely implemented due to missing incentives for data authors to prepare datasets and software code, data citation is expected to facilitate rewarding data sharers, provide detailed attribution and enhance collaboration in scholarly communication.

White (1982) called for the needs of citing datasets in the social science context from the early 1980s. The citation analytic approach is important for data citation. Citation merits study because it represents one of the major rewards and opportunities for formal recognition for authors (e.g., data sharers) within the scholarly community. Data citation involves reference to the data themselves (rather than to publications that share data) in order to give attribution, to facilitate access (CODATA-ICSTI Task Group on Data Citation Practices, 2013), and to promote direct and unambiguous reference to datasets in a study. The availability of datasets may be reported in data journals such as Nature’s *Scientific Data*.

Data citation has been more actively studied in the realm of data sharing than that of data reuse, perhaps owing to the labor-intensive processes involved. Thus, for example, data collection processes include the manual review of full texts, references, and supplementary datasets (necessitating, e.g., the opening of supplementary datasets in a file format such as .pdf or .doc). Moreover, in the absence of sufficient domain knowledge, data collection for reuse has a high potential for inaccuracy, since the hard sciences and/or engineering demand expert domain knowledge in order to identify reused data residing within the full text of an article.

A persistent identifier should be assigned at the time of data publication, and the low frequency with which such identifiers as a DOI or researcher ID (e.g., an ORCID ID) actually are assigned creates significant challenges in the identification of significant factors relating to data citation practices.

Previous literature has explored such topics as data citation principles, standardization, peer review for data publication, practices, infrastructure, metadata elements associated with a dataset (rather than embedded within it; e.g., provenance metadata rather than descriptive metadata), DOIs (digital object identifiers, both unique and persistent, that include a time-stamp and version history), technical infrastructure, flexibility for interoperability across communities, policies regarding repositories and data journals, data management practices best suited to research, the high incidence of self-citation, citation protocols, altmetrics, and linked data (CODATA-ICSTI Task Group on Data Citation Practices, 2013; Lawrence, Jones, Mattews, Pepler, & Callaghan, 2011).

These previous studies of data citation, while having their limitations, provide valuable insights. All the same, the primary focus has been on individual disciplines rather than on the impact of data citation across such disciplines as science, technology, and engineering. Data sharing varies within each discipline (Tenopir et al., 2011), which means that the impact of data citation on data sharing cannot be fully appreciated without considering disciplinary factors. A gap thus remains in the literature with regard to the impact of data citation within and across the diverse science, technology, and engineering disciplines.

Previous studies, then, have used descriptive statistics (e.g., distributions) regarding the history of citation, but analysis has not actively been studied for data citation. Also, previous literature has focused on the practices of data citation from the DCI or on a single data repository (e.g., CIPSR or Dryad) or has used the history of citation. Examples of areas that have been studied include (a) journal policies regarding metadata (e.g., data descriptors regarding dataset stories of high-profile journals), (b) citation practices within full texts (e.g., accession numbers provided in the full text of articles), and (c) manual review by looking into practices regarding references, full texts, and rewards and acknowledgements among journals. For datasets in data citation, research has focused on (a) the type of datasets/data study, (b) practices of the DCI, and (c) restrictions (e.g., restricted/limited/unrestricted datasets stored in data repositories).

Data journals can impact data citation; for a data journal article is not a traditional paper. Rather, data journals provide quality data (e.g., peer-reviewed research data) that may be used by or of interest to others and that includes the main metadata elements that map to the concept of citation (CODATA-ICSTI Task Group on Data Citation Practices, 2013). In current practice, data journals do not (yet) require peer-reviewed data as a journal policy. Data journals have emerged as an alternative to the direct data citation (Belter, 2014). Examples of data journals include S*cientific Data, PLoS, Data Science Journal, BMC Research Notes, Journal of Open Archaeology Data* and *Biomedical Data Journal*. Data journals may promote data citation through their policies, for example by requiring that datasets be included in the reference list of a paper or that DataCite recommendations be followed. A few studies have dealt with data journals. Using a survey method, Candela, Castelli, Manghi, and Tani (2015) have looked into 100 existing data journals in terms of dataset description, availability, citation, quality, and open access in order to identify ways to expand and strengthen the data journal approach that will increase access to and exploitation of datasets. Thus, “first principles” for data citation have been identified by CODATA-ICSTI Task Group on Data Citation Practices (2013, p. CIDCR6):

- Status of data: Data citation should be accorded the same importance in the scholarly record as the citation of other objects.
- Attribution: Citations should facilitate giving scholarly credit and legal attribution to all parties responsible for the data.
- Persistence: Citations should be as durable as the cited objects.
- Access: Citations should facilitate access both to the data themselves and to such associated metadata and documentation as are necessary for both humans and machines to make informed use of the referenced data.
- Discovery: Citations should support the discovery of data and their documentation.
- Provenance: Citations should facilitate establishment of the provenance of data.
- Granularity: Citations should support the finest-grained description necessary to identify the data.
- Verifiability: Citations should contain information sufficient to identify the data unambiguously.
- Metadata standards: Citations should employ widely accepted metadata standards.
- Flexibility: Citation methods should be sufficiently flexible to accommodate variant practices among communities but should not differ so much that they compromise interoperability of data across communities.

Current practices in data citation do not give due credit by linking bibliographic references to published research data because published research data tend to be regarded as supplementary material (CODATA-ICSTI Task Group on Data Citation Practices, 2013; Park & Wolfram, 2017). It is argued, however, that data citation should accompany such published works as articles in a references section in order to give due credit to data sharers (e.g., data authors). Tenopir and colleagues (2011) found that 91.7% of them somewhat agreed with the importance of their shared data being cited if their shared data are reused by other researchers. Data sharing and reuse across multiple disciplines thus remains relatively unexplored from the perspective of data citation from scholarly databases, data journals, or data repositories. In practice, data citation is (still) far from common (Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2016). Data citation should be formally cited in bibliographic references section to give due credit to data sharers as noted by Park and Wolfram. The practices in biomedical fields show that informal data citation, in which data citation is mentioned in passing in the main texts or out of references is more commonly found than formal data citation, in which data citation is in the references section (Park, You, & Wolfram, 2017)

The DCI, which was launched in 2012 by Thomson Reuters and was sold in 2016 to Clarivate Analytics, currently provides data citation indexing as a subscription-based service. The DCI provides a single access point to over 350 data repositories worldwide and thus to over 7.4 million records across multiple disciplines (Clarivate Analytics, 2018). The DCI divided its records into 4 major categories. Major 4 categories are dataset, software, data study and repository. The records by the DCI treats datasets in a similar way as journal articles or other document types such as conference proceedings or books in the bibliographic WoS databases. The citation records of the DCI are connected to related literature indexed in the WoS database. An advantage of the DCI is that indexers of the Clarivate Analytics WoS, and even Elsevier's Scopus, can detect and track data citation. The DCI has been used to examine why research data are cited in genetics and heredity (Park & Wolfram, 2017), in biomedical fields (Park, You, & Wolfram, 2018) and in the humanities (Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2016).

The uniform direct citation of data curated in persistent data repositories has been emphasized with regard to the reproducibility and reusability of research outcomes because "a foundational element of reproducibility and reusability is the open and persistently available presentation of research data" (Starr et al., 2015). As discussed above, a permanent and persistent data identifier (e.g., a DOI) at the time of data publication may be important for direct citation and ease of accessibility in data citation because a DOI is machine-readable and therefore provides access to cited data and its associated metadata. However, current practices are such that data citation includes only a low percentage of persistent identifiers (e.g., DOI or ORCID) compared with regular citation. Access to data repositories (i.e., open access data repositories), whether unrestricted, limited, or restricted, thus needs to be studied in the context of data sharing and the potential future reuse of data.

Direct citer-based analysis has been conducted in data citation research with the comparison of research data citation and citing articles in Genetics and Heredity in order to identify self-citation and recitation (Park & Wolfram, 2017). Applying direct citer-based analysis across multiple disciplines (i.e., interdisciplinarity) remains a promising avenue to be explored as a means to measure author research impact in the context of high rates of self-citation (i.e., of authors citing themselves). The same authors tend to use the same shared research data repeatedly, potentially indicating a high rate of self-citation. Relatively greater numbers of publications cited by a citing author indicate relatively greater influence of the cited author on the citing author. Direct citer-based analysis has been discussed recently (Ajiferuke, Lu, & Wolfram, 2010), though citer-based analysis from a more general perspective has been the subject of ongoing study.

The citation rate is associated with research data sharing because it provides a detailed description of data (Piwowar & Chapman, 2010; Piwowar, Day, & Fridsma, 2007). Data sharing and future data reuse may be increased in the case of secondary analysis. Piwowar and Chapman studied 397 gene expression microarray datasets published in 2007 in 20 different journals and report that investigators are more likely to share their raw datasets publicly on the Internet when their research are published in high-profile journals and when the first and last authors have had high-impact careers. Publishers of lower-impact journals do not enforce their data-sharing policies rigorously. Piwowar, Day and Fridsma determined that 69% more citations occurred between microarray clinical publications and their associated data sharing when data were publicly available. Their examination of citation history used multivariate linear regression to reveal that public data sharing is significantly associated with increased citation rates, independent of the journal impact factor, publication date, and author's country of origin.

The major institutional bases of disciplines at the levels of college/school, department, and lab need to be identified based on author affiliation. Author affiliation data can be found in the headers of an article and in the acknowledgements. The presentation of supplementary material in the relevant location may facilitate automatic or machine-actionable data citation with bidirectional links between articles (e.g., in the case of a data journal), associated datasets, and data repositories. Examples include (a) supplementary material inserted at the point of reference/citation, (b) placing the material in the proper context, (c) making supplementary material easier for readers to find, and (d) locating supplementary material initially in a closed text-box.

Metadata plays an essential role to trace, access and effectively use research data. Research data must be accompanied by basic descriptive metadata. The Dublin Core (DC) is closely aligned with the mandatory fields in the DCI because the parts of 15 elements of the DCI metadata such as creator, title, publication year and identifiers are widely used as mandatory fields to data preservation. Although this alignment to the DC allows interoperability across different platforms, ambiguity also increases for the detailed study of data metrics for research evaluation. Including rich metadata such as provenance metadata, rights metadata (e.g., license information) and technical metadata (e.g., file size) would facilitate to actual access to shared research data and allow for the description of discipline specific research data as well. Adequate information of metadata for data reuse demands researchers to fill out the form of fields to characterize shared data. Data Documentation Initiative (DDI) provides extensive guidelines of metadata for many forms of human subject research by developing a data model for qualitative data. The major challenges of having comprehensive metadata in order to provide adequate information for data reuse are how to explain the particularities of specific portions of research data.

Metadata formats for data citation are emphasized in earlier literature (Borgman, 2012). However, metadata in data citation is inconsistent at present. The literature has noted that the consistency, quality and sustainability of metadata in research data need to be studied (CODATA-ICSTI Task Group on Data Citation Practices, 2013; Helbig, Hausstein, & Toepfer, 2015; Starr et al., 2015). Quality control is mainly mentioned for reliable data reuse for reproducibility. In maintaining metadata, sustainability is another concern (Helbig, Hausstein, & Toepfer, 2015). For research data, the Dublin Core Metadata Element Set has been widely used in order to develop application profile in a given context, for instance, Dryad Application Profile (Ball, 2009; Diamantopoulos, Sgouropoulou, Kastrantas, & Manouselis, 2011). However, due to the DC's relatively flat structure, the complex relationships of software or research datasets confronted challenges (Lagoze, 2000). The DataCite Metadata Schema is one of the approaches to overcome these challenges because the DataCite Metadata Schema can describe the relationships between two datasets (DataCite Metadata Working Group, 2016; Star & Gastl, 2011). The DataCite Metadata Working Group (2015) released the DataCite Metadata Schema as the core metadata properties to consistently identify a resource for data citation and retrieval with the recommended instructions. Metadata in data citation needs to be taken into account for the administrative or methodological metadata rather than descriptive metadata (Star & Gastl, 2011). For instance, Chao (2015) examined methods metadata in soil science such as common methods-related elements of articles. A previous study (Canham & Ohmann, 2016) examined metadata scheme in clinical research and proposed elements for metadata scheme in clinical research data into three: mandatory, recommended and optional. Mandatory elements in clinical research include source study title, DOI, title, creators, creation year, resource type in general, publisher, access type, access details, access contact and resources. Recommended elements in clinical research include study identifier, study topics, version, resource type, description, language and other hosting institutions. Optional elements in clinical research include object other identifiers, object additional titles, contributors, dates, subjects and rights.

Disciplinary metadata standards (Digital Curation Center, 2018) are in practice because each discipline will process their research data differently and will use different vocabularies to describe research data. Examples are Darwin Core, Ecological Metadata Language (EMI) and Genome Metadata in biology. In earth sciences, examples of disciplinary metadata in use are Astronomy Visualization Metadata (AVM), Climate and Forecast (CF) Metadata Conventions and Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (FGDC/CSDGM).

General metadata for research data are also in use that includes DataCite Metadata Schema, Data Catalog Vocabulary (DCAT), DC and Repository-Developed Metadata Schemas.

In order to identify relevant research methods in data citation, Table 2 provides a comparison of various methodologies that have been used to study data citation. Some of these works have used mixed methods approaches, although relatively few studies have been conducted on the citation of peer-reviewed data. For example, Fear (2013) has combined quantitative methods, including logic regression and bivariate analysis, with such qualitative methods as content analysis and interviews. Park and Wolfram (2017) combined quantitative methods, including citer-based analysis and citation-based analysis, with such qualitative methods as content analysis (e.g., manual assessment). Domains that have been studied with regard to data citation include genetics and heredity (Park & Wolfram, 2017), and the social sciences (Fear, 2013).

As seen in Table 2, these methods, rather than being specific to data citation, are associated with data sharing and reuse. To be specific, previous studies have mainly relied on: (1) quantitative methods involving (i) surveys with closed-ended questions and (ii) regression for quantitative analysis; (2) qualitative content analysis; and (3) mixed methods combining (i) surveys with closed-ended questions and (ii) qualitative content analysis in the case of mixed methods approaches. Informetrics approaches have been studied only relatively recently (Fear, 2013; Peters, Kraker, Lex, Gumpenberger, & Gorraiz, 2016; Park & Wolfram, 2017; Piwowar & Chapman, 2010).

Regarding qualitative methods, content analysis and interviews have been the main instruments used in the study of data citation. Thus, to return to a previous example, Fear (2013) has employed both approaches in order to explore significant factors that improve data reuse associated with ICPSR repository in the social sciences. Park and Wolfram (2017) employed content analysis with the manual assessment of references, main text, acknowledgemenet, funding information, supplementary information, and author information to examine hidden or embedded data citations regarding data sharing and reuse in Genetics and Heredity.

Informetrics-based methods provide promising approaches to the study of data citation. Previous researchers have explored quantitatively phenomena such as: (1) citation analysis based on citation history including direct citation and co-citation (Piwowar, 2010), and self-citation (He & Nahar, 2016); (2) allied analysis, including co-authorship analysis (Fear, 2013) with the aim being to examine collaboration and natural language processing (NLP) (Piwowar, 2010) or co-word analysis in studying text and language; and (3) the use of citer-based analysis to identify self-citation (Park & Wolfram, 2017). In the case of collaboration, citer-based analysis may represent a remedy for self-citation (Ajiferuke, Lu, & Wolfram, 2010; Lu, Ajiferuke, & Wolfram, 2014), and communication detection using map equations may be approached based on information flow (Bohlin, Edler, Lancichinetti, & Rosvall, 2014) in order to identify dynamic areas. Quantitative surveys have been used to explore significant factors affecting data citation in the sciences (Candela, Castelli, Manghi, & Tani, 2015; Swauger & Vision, 2015), naturally making greater use of closed- rather than open-ended questions (Curty, 2015; Swauger & Vision, 2015).

Table 2 Summary of prior studies on methodologies used to study data citation

| **type** | **quantitative methods** | **qualitative methods** | **sources** |
|---|---|---|---|
| data citation | survey | - | (Candela, Castelli, Manghi, & Tani, 2015) |
| | - | literature review | (CODATA-ICSTI Task Group on Data Citation |

| | | | |
|---|---|---|---|
| | | | Practices, 2013; Silvello, 2018) |
| | citation analysis, descriptive analysis | - | (Peters, Kraker, Lex, Gumpenberger, & Gorraiz, 2016) |
| data citation / sharing | survey | content analysis | (Swauger & Vision, 2015) |
| | citation analysis, regression, exploratory factor analysis, univariate/multivariate | - | (Piwowar, 2010) |
| | multivariate logic regression analysis | - | (Piwowar & Chapman, 2010) |
| | multivariate linear regressions, correlation | manual review (of citation context) | (Piwowar & Vision, 2013) |
| data citation / reuse | citation analysis | - | (He & Nahar, 2016) |
| | bibliometric analysis logistic regression, bivariate analysis | content analysis, interview | (Fear, 2013) |
| data citation | citation analysis, citer-based analysis, descriptive analysis | content analysis | (Park & Wolfram, 2017) |

| | | | |
|---|---|---|---|
| / reuse/ sharing | descriptive analysis | content analysis | (Park, You, & Wolfram, 2018) |

Previous literature has addressed a variety of related topics, including data citation principles, standardization, peer review for data publication, practices, infrastructure, metadata elements associated with a dataset rather than embedded within it (e.g., provenance metadata rather than descriptive metadata), DOIs (digital object identifiers, both unique and persistent, which include a time-stamp and version history), technical infrastructure, quality control in data reuse, flexibility for interoperability across communities, policies regarding repositories and data journals, the best data management practices for research, the high incidence of self-citation, citation protocols, altmetrics, and linked data (CODATA-ICSTI Task Group on Data Citation Practices, 2013; Lawrence, Jones, Mattews, Pepler, & Callaghan, 2011).

Persistent identifiers are important for data citation because data citation needs a unique and persistent identifier for reusers to obtain the latest available version and format of the resource. "A persistent identifier enables unambiguous referencing, cross-referencing, authentication and validation… provides a basis for practices such as citation counting in career merit reviews" (CODATA-ICSTI Task Group on Data Citation Practices, 2013, p. 15). As discussed above, a persistent identifier should be assigned at the time of data publication, and the low frequency with which such identifiers as a DOI or researcher ID (e.g., an ORCID) actually are assigned creates significant challenges in the identification of factors that are important for data citation practices.

#### 2.4.3.1.Data Repository Impact

Data repositories are storage and publication platforms where research data are disseminated as research output. The advantage of sharing research data in repositories includes an easier and more standardized data transfer between journals and data repositories. Types of repositories include general-purpose repositories, discipline-specific repositories and institutional repositories. Examples of repositories include Zenodo (https://zenodo.org/), figshare (https://figshare.com/), GenBank (https://www.ncbi.nlm.nih.gov/genbank/), PANGAEA (https://www.pangaea.de/), Harvard Dataverse (https://dataverse.harvard.edu/) and UniProtKB (https://www.uniprot.org/). Repositories such as Dryad and Harvard Dataverse generate a data citation directly.

Data repositories are important for providing the raw data used by DCI meaning repositories play an essential role in data citation for scientific knowledge dissemination by providing metadata, persistent access (e.g., DOI), stewardship and data discovery to find research data. In 2010, the announcement of the *Journal of Neuroscience* stopped publishing supplementary materials and promoted disciplinary repositories (Maunsell, 2010), indicating journal publishers' recognition of the importance of data repositories. Journal publishers suggest or recommend data citation in a domain-specific list of acceptable repositories. For instance, Nature publishing group (2018) provides the recommended data repositories by each discipline for the data journal called *Scientific Data*.

To measure data repository impact, data repositories need to provide research data in forms to be citable and descriptions to be understandable for data sharers and reusers. However, the citations of data repositories are not common. For instance, 43 repositories in the DCI did not receive any citations (Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2016).

Although there has been a focus on the general-purpose or discipline-specific repositories, institutional repositories also have been examined. Fan (2015) examined 19 institutional repositories affiliated with the Chinese Academy of Sciences with the webometric indicators of their home institutions, especially for the citation rate of papers in home institutions. Fan found that institutional repositories can improve the visibility of their home institutions and the web presence. Also, if the institutional repositories are open access, their home institutions received more web visibility and presences.

#### 2.4.3.2.Data Citation Impact

Data citation is important for data sharing. As just noted, there is a 69% increase in the citation rate of published research when detailed information is provided for shared data (Piwowar, Day, & Fridsma, 2007; Piwowar, 2010) and "independently of journal impact factor, date of publication, and author country of origin using linear regression" (Piwowar., 2010, p. 14). According to one estimate, however, 43% of repositories received no citations (Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2016) and 61% of datasets stored in the ICPSR repositories did not provide any type of citation to datasets (Mooney, 2011). In a recent survey, 91.7% of researchers somewhat agreed that data citation is important when their data are reused (Tenopir et al., 2011), and 95% agreed that it is "fair to use other people's data if there is formal citation of the data providers and/or funding agencies in all disseminated work making use of the data" (Tenopir et al., 2011, p. 10). The mechanism of data citation and publication, involving citable, easily discoverable, and reusable research output, provides an incentive for researchers to document and archive data appropriately (Callaghan et al., 2012).

In the social sciences, data citation is not new. In the late 1970s, recommendations to reference machine-readable data files (MRDF) were published by Dodd (1979). Dodd suggested guidelines for the data citation in social science regarding the format of references to MRDFs. White (1982) mentioned the importance of formal citation apart from main text in social sciences. White mentioned that "[a]n argument by no means new is that social scientists who work with machine-readable data files (MRDF) should cite them in their writings, with formal references set apart from main text, just as they now do books, papers and reports (p. 467)". White found that data citation is highly incomplete and inconsistent and demands considerably further studies with the examination of three sets of data files in the WoS's SSCI. In the ICPSR, one of the largest data repositories in social science, 61% of articles among 49 journal articles did not formally cite articles (Mooney, 2011). However, the form of informal data citation, which was mentioned in passing by the dataset title, was widely found in social science (Mooney & Newton, 2012). Also, confidentiality and anonymity are crucial requirements for qualitative data sharing. Protecting confidentiality and data with sensitive information are the most frequently mentioned barriers for the qualitative data sharing and preservation by researchers (Cliggett, 2013).

Sustainable persistent methods for data identifications such as DOIs are needed for data citation. Examples of existing global identifiers are the DOI, ORCID, Uniform Resource Name (URN), the Life-Science Identifier (LSID) and Research Resource Identifier (RRID). However, in astronomy, over 40% of data linked via URLs in the astronomical literature are broken in a decade of publication. DOI can minimize this problem because when the associated Uniform Resource Locator (URL) change, the registration agency can update the DOIs via an Application programming interface (API; Pepe, Goodman, Muench, Crosas, & Erdmann, 2014). The persistent and sustainable identifier of the research data resolves to a correct landing page must support multiple levels of granularity.

#### 2.4.3.3.Data Sharing Impact

The impact of data sharing has been studied from the perspective of a single rather than multiple disciplines. For example, Piwowar (2010) studied data citation of the data sharing in biomedicine. The impact of data sharing impact has revealed that the sharing of data publicly increases the citation rate of publications (Piwowar, Day, & Fridsma, 2007; Piwowar & Vision, 2013). Researchers are more inclined to share their research data if researchers receive credit (Borgman, 2012).

Research data are usually considered as the primary data source in conducting scholarly research. Scientists' data sharing behaviors (Kim & Stanton, 2015), perceptions (Tenopir et al., 2011), and cultures have been actively studied, with a focus on the barriers to data sharing in public repositories. Relatively recently, data repositories themselves have begun to be studied. A review of previous literature reveals that data journals are not yet the focus of active research. The benefits of sharing research data include the validation of findings and the potential future reuse of shared data.

Domains studied with regard to data sharing impact have been actively studied in the field of biomedical fields. For instance, the data sharing impact has been examined in such fields as biomedical microarray (Piwowar & Chapman, 2010) genetics and heredity (Park & Wolfram, 2017) and biomedical fields (Park, You, & Wolfram, 2018). Again, individual domains, rather than multiple disciplines, have been actively studied regarding data sharing impact (Park & Wolfram, 2017; Park, You, & Wolfram, 2018; Piwowar, Day, & Fridsma, 2007; Piwowar & Chapman, 2010), at least until recently.

Factors influencing data sharing and withholding can be usefully categorized into three groups: institutional factors (e.g., a funding agency's policy, journal requirements, and contracts with industry sponsors), resource factors (i.e., metadata and data repositories), and individual factors (e.g., personal characteristics, perceived benefit, perceived effort, and perceived risk). Other organizational and environmental factors have also been identified as significantly influencing scientists' data sharing and withholding (Kim, 2013). Journals' data-sharing policies do not necessarily induce authors to make their research datasets accessible to independent investigators (Savage & Vickers, 2009).

#### 2.4.3.4.Data Reuse Impact

Data reuse impact has not been actively studied. A challenge impeding the study of data reuse is that researchers mainly need to use manual methods by scanning research literature in order to identify reused data in scientific publications. Several studies have proposed methods for streamlining the data reuse impact. Abstracts rather than full-text in biomedical fields by using NLP techniques have been applied due to the free and standard format of abstracts (Lin, 2009) although more information is contained in the full-text of literature. By using indicating terms of data sharing and reuse (Park & Wolfram 2017), data reuse impact has been examined with semi-automatic ways by using automatic text searching techniques and manual human judgments in biomedical fields (Park, You, Wolfram, 2018). Main-text (i.e., full text) of high-profile articles mainly contains research data reuse (Park & Wolfram, 2017), which prevents the automatic indexing of data citation. Currently, there does not seem to be appropriate tools to measure data reuse automatically.

The impact of data reuse has not been actively studied in the context of multiple disciplines and multiple data repositories. Fear’s (2013) study looks at the social sciences by analyzing only a single data repository, the ICPSR, using a mixed-methods approach. Multiple data repositories in the DCI (i.e. over 350 repositories) in a single discipline are examined from the informetrics approaches (Park & Wolfram, 2017). A data paper as an incentive mechanism has the potential to advance data publishing to the level of scholarly publishing and to lead to a significant increase in efficiency, at least in the field of biodiversity science (Vishwas & Lyubomir, 2011). Data peer review has been shown to improve data quality, though there is no formally established or recognized process (Parsons, Ruth, & Minster, 2010).

## 2.5. Software Sharing, Reuse and Citation

### 2.5.1. Software Sharing

Studies of scholarly communication have also investigated software sharing. This is a continuous process that merely begins with the initial sharing, for software can be updated as new versions (e.g., Version 1, Version 1.1 or Version 2, and so on) are disseminated in order to correct bugs or in response to users’ requests for more advanced functions. Software can be shared in a variety of ways, such as among local teams (e.g., code shared through a laboratory research team’s local server), by means of personal websites, or through repositories used in scholarly communication, such as the Comprehensive R Archive Network (CRAN) or Zenodo. Software sharing faces numerous impediments, though; thus, for example, Howison and Herbsleb (2011) found it to be costlier and more complicated than the sharing of data or the circulation of traditional publications.

Software is frequently mentioned in scholarly communication involving scientific publications, as Li and Yan (2018) found with respect to R packages being referred to widely across PLoS papers. Especially significant for the present study, Pan, Yan, & Hua (2016) reported that open software was mentioned more often than proprietary software in the full texts of *PLoS ONE* articles published in 2014. Findings such as these indicate, then, that open software sharing increases the attribution of scholarly credit for those who share software.

### 2.5.2. Software Reuse

The reuse of software, which was first discussed half a century ago by McIlroy (1968), remains a major concern for the software engineering community. Reuse can minimize the time required to create software, contribute to the stability of systems thanks to reliance on previously tested and created components (Krueger, 1992), and improve the overall quality of software. Software, source code, and online programming resources are widely accessible in the context of open source projects. Not surprisingly, the increased accessibility of software for reuse is changing the ways in which programmers write their program languages, as they often opt to copy and paste existent program code from various sources. So, it is that, according to one estimate, fully half of the code being created for production reuses code from previous programs (Mockus, 2007). This situation creates challenges for programmers, who spend considerable time searching for appropriate pieces of existing software when specific needs arise. Software reuse can thus be considered one form of data reuse and can usefully be distinguished as either architecture, design, or program reuse (Aziz & North, 2007).

Various factors determine the success of software reuse, which can be defined as "the systematic practice of developing software from a stock of building blocks, so that similarities in requirements and/or architecture between applications can be exploited to achieve substantial benefits in productivity, quality, and business performance" (Morisio, Ezran, & Tully, 2002, p. 341). Larger-scale reuse, then, is supported by smaller scale reuse (Henry & Faller, 1995). The success of large-grained software reuse within an organization depends on such factors as trust and organizational culture (Witman, 2007). There are also a number of barriers to higher-level software reuse, whether conceptual (e.g., failure to understand the basic elements of reuse), technological (e.g., lack of common standards across or within organizations; poor practices with regard to software metrics), infrastructural (e.g., obsolete supporting infrastructure), managerial (e.g., lack of consensus regarding common standards across diverse projects), or cultural (e.g., disincentives to efficient reuse within large development teams) (Bassett, 1997).

### 2.5.3. Software Citation

The importance of software in scientific research can hardly be overstated; thus, according to a recent report by the National Postdoc Association, 95% of postdoctoral researchers use software, and 63% could not do their work without it (Nangia & Katz, 2017). Nevertheless, it is only recently that research software citation has been actively studied from the perspectives of software sharing and reuse (Hong, Hole, & Moore, 2013; Li, Yan, & Feng, 2017; Pan, Yan, Cui, & Hua, 2018). Technically, software is a form of data (Marcus & Menzies, 2010) and, if curated in data, it can be given due scholarly credit (Lynch, 2014). Software is, of course, different from data, in particular because it is executable as a creative work (Katz et al., 2016). On the other hand, like data, it has not traditionally been included in journal publications.

Previous studies observed inconsistent software citation (Howison & Bullard, 2016; Katz & Smith, 2015; Li, Yan, & Feng, 2017). Howison and Bullard found that many software entities do not provide consistent information regarding the form of citation. Li, Yan and Feng found that R packages in published articles showed inconsistent practices from formal citation to informal citation. In biology, a random sample of 90 articles showed several different ways of software mentioning such as main text, URLs in footnotes, different kinds of mentions in the references section (Howison & Bullard, 2016), brining difficulties of formal software citation. Due to these inconsistencies, the development of proper software citation entities in published research outputs has been recognized by the FORCE 11 Software Citation Working Group (Katz & Smith, 2015). The software citation principles include importance, credit and attribution, unique identification, persistence, accessibility and specificity (Smith, Katz, Niemeyer & FORCE11 Software Citation Working Group, 2016). More specifically,

- Importance: Software should be regarded as a citable product of research with the same importance such as journal publications in scholarly communication.
- Credit and attribution: Software citation should facilitate giving rewards to all contributors of the software in scholarly communication.
- Unique identification: Software citation should have a machine actionable, unique and interoperable identification.
- Persistence: Unique identification and metadata of software and its disposition should be persistent.
- Accessibility: Software citation should facilitate software accessibility by making the information to the referenced software usage.

- Specificity: Software citation should facilitate the identification, access and version specification of software with specific identification needed such as version number and revision numbers.

Informal software citation has been more common than formal software citation, at least until recently. Its prevalence prevents software sharers from receiving due scholarly credit for publishing in highly-impact journals (Poist, 2015). The fact that 97.7% of BMC Bioinformatics papers mention software and databases in passing is indicative of the high rate of informal software citation. Howison and Bullard (2016), based on the aforementioned survey of 90 biology articles, reported that 31% of informal software citation took the form of passing mentions in the text, while 44% provided formal citation. Likewise, only 13% of some 1,000 publications analyzed in another study specifically mentioned the software used in generating the research outcomes, and only 50% of the publications acknowledged individuals personally, which is another type of informal software citation (Weber & Thomer, 2014). A more recent study found formal citations to be common in *PLoS* journals when official citation instructions are provided (Li, Yan, & Feng, 2017).

Formal citation is clearly important for sharers of research software, and journals indexed by such scholarly databases as WoS and Scopus provide a venue for it. Although researchers' sharing of code with the public at no cost is motivated by the desire to enhance their own academic reputations and to receive credit for their work (Poist, 2015), formal software citation remains relatively rare, as is the case, for instance, in the geosciences (Reichman, Jones, & Schildhauer, 2011). The form that formal software citation takes can also have a significant impact on its pervasiveness, as demonstrated by the Text Retrieval Conferences (Rowe, Wood, Link, & Simoni, 2010).

Software journals can provide a venue for formal indexing by scholarly databases, such as WoS and Scopus, and thereby for formal software citation, as well as for publication of the software itself. These journals publish articles focused entirely on research software and attribute formal scholarly credit. According to Soito and Hwang (2016), papers that describe software compel researchers to identify the specific code used. Examples of software journals include the *Journal of Open Source Software* and *Journal of Open Research Software*. Domain-specific software journals, such as *Computer Science Communication* and *Bioinformatics*, have traditionally accepted research software submission from authors.

Software measurement ontology and elaboration of a multi-level metadata framework are two recent initiatives addressing software citation. The former involves efforts by researchers to approach citation from the perspective of software development (García et al., 2006). As for the multi-level metadata framework, it has been developed as a means to describe the reusability of software by developers (Hong, 2014). Li, Lin, and Greenberg (2016) analyzed current practices relating to inconsistent descriptive metadata elements and the types of software reuse in 400 papers in the field of material sciences, looking specifically at a popular piece of simulation software, Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) and concluded that inconsistent metadata associated with research software could limit accessibility to it and thus, ultimately, citation of it. Not surprisingly, available software metadata for content specifications vary across communities. Ontosoft (Gilbert, 2015), for instance, is a community software commonly used in the geosciences. Existing software metadata can be language-specific, examples being R package descriptions (Bechhofer et al., 2013) and Python packages (Ward & Baxter, 2016). Terms and classes are also defined at schema.org (https://schema.org/SoftwareApplication).

## 2.6. Summary

This chapter has reviewed the relevant literature regarding informetrics and the current state of data citation in scholarly communication and in the process has outlined the overall environment and context of scholarly measurement. Research data citation is, then, attracting increasing attention, but relatively little work has been done on the topic. As has been seen, the research that has been conducted suffers from certain limitations regarding generalizability and the lack of a methodological framework. The present study was designed to help fill in some of these gaps. The various types of metric studies of scientific communication were reviewed (Section 2.2) because of their applicability to the study of data citation, as was the concept of open science as it relates to data sharing (Section 2.3), discussion of which included a comprehensive review of concepts and research relating to data and software citation. Next, research on the measurement of scholarly impact was surveyed with a focus on data sharing and reuse (Section 2.4). Lastly, work on research software citation was approached from the perspectives of software sharing and reuse (Section 2.5).

# Chapter 3 METHODOLOGY

This chapter outlines methodological frameworks for the study of data citation in the context of data sharing and data reuse. This discussion is important because a framework for useful information relating to data citation has not yet been developed because it is a relatively new phenomenon and the study of data journals and citation is in its infancy. This being the case, informetric methods and text searching provide useful analytical tools for exploring data citation. The mixed methods approach employed here combined quantitative informetrics and qualitative semi-automatic content assessment. One contribution of this dissertation is thus the establishment of a methodological framework, specifically a refined research model that takes into account key previous studies of data citation, sharing, and reuse, in particular those that have identified groups of factors relating to these activities. The data analysis methods used in this study were primarily quantitative, but a qualitative component was included in the evaluation of data reuse.

## 3.1. Introduction

The combination or mixing of quantitative and qualitative approaches can provide comprehensive perspectives for the study of complex social phenomena (Creswell & Clark, 2011). Quantitative methods have the capacity to yield generalizable results when representative samples are used. Qualitative methods, on the other hand, are called for when addressing complex questions that cannot be answered with quantitative methods and can serve to provide a comprehensive and in-depth examination of phenomena. In my research, a mixed method approach allowed me to answer my research questions to more clearly examine and understand the phenomenon of research data citation based on sharing and reuse in STEM fields. Quantitative approaches are used to capture the phenomenon of data citation effectively. Qualitative approaches, based on the manual assessment of data reuse, provided a rich context for identifying data citation.

Table 3 summarizes the primary concepts that, according to prior studies, may be associated with research data citation in scholarly communication. Specifically, the following concepts that were associated with data citation were conceptualized based on data sharing, data type, self-citation and discipline. Taking each of these concepts in turn, sharing makes research data citable and reusable (Helbig, Hausstein, & Toepfer, 2015) and increases the citation rate for articles (Dranchen, Ellegaard, Larsen, & Dorch, 2016; Gordon et al., 2016; Helbig, Hausstein, & Toepfer, 2015; Piwowar & Vision, 2013; Piwowar, Day, & Fridsma, 2007). Regarding types of data, certain sources, such as surveys, aggregated data, and sequence data are more often cited than others (Peters, Kraker, Lex, Gumpenberger, & Gorraiz, 2015). Regarding disciplines, each has its own distinct data sharing practices owing to its unique citation behavior (Helbig, Hausstein, & Toepfer, 2015; Torres-Salinas, Jiménez-Contreras, & Robinson-García, 2014); the actual rate of data sharing also varies within scientific communities (Tenopir et al., 2011). Lastly, regarding self-citation, the same authors tend to use the same shared research data (Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2016), and self-citation and author self-citation are prevalent in research data citation (Park & Wolfram, 2017).

Table 3 Concepts that are associated with data citation

| concepts | justification | sources |
|---|---|---|
| **data sharing** | Researchers are more inclined to share their research data when they receive credit and the lack of a reward system discourages researchers' data sharing. | (Borgman, 2012; Costas, Meijer, Zahedi, & Wouters, 2013) |
| | Sharing makes research data citable and reusable for secondary research. | (Helbig, Hausstein, & Toepfer, 2015) |
| | Articles with shared research data have higher citation rates than those without, and therefore greater impact. | (Dranchen, Ellegaard, Larsen, & Dorch, 2016; Gordon et al., 2016; Helbig, Hausstein, & Toepfer, 2015; Piwowar & Vision, 2013; Piwowar, Day, & Fridsma, 2007) |
| **data type** | Qualitative data tend to be rarely shared for reuse | (Faniel & Jacobsen, 2010; Wallis, Rolando, & Borgman, 2013; Yoon, 2014) |
| | The experimental data are mostly reused by researchers. | (He & Nahar, 2016; Zhao, Yan, & Li, 2018) |

| | | |
|---|---|---|
| | Certain types of data, such as surveys and aggregated and sequence data, are more often cited and receive higher altmetrics scores. | (Belter, 2014; Peters, Kraker, Lex, Gumpenberger, & Gorraiz, 2015) |
| **self-citation** | 84 percent of scientific data citations are self-citing in Dryad repository. | (He & Nahar, 2016) |
| | Self-citation and author self-citation are prevalent in research data citation. | (Park & Wolfram, 2017) |
| | The same authors tend to use the same shared data. | (Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2016) |
| **discipline** | Each discipline has its own distinct data sharing practices. | (Helbig, Hausstein, & Toepfer, 2015; Torres-Salinas, Jiménez-Contreras, & Robinson-García, 2014) |
| | Depending on the subject category in the DCI, data sharing practices are very diverse. | (Park & Wolfram, 2017) |
| | Within scientific communities, the actual rate of data sharing varies from discipline to discipline | (Tenopir et al., 2011) |

## 3.2. Data Collection

This research used Clarivate Analytics' Web of Science (WoS) as a data source rather than Elsevier's Scopus because the recorded citations in the former cover a longer period than the latter by an average of 30%, though the historical record varies somewhat across disciplines (Leydesdorff, de Moya-Anegón, & Guerrero-Bote, 2010). Scopus does include more journals in the social sciences and the humanities fields than WoS, but this advantage was not relevant to the present study given the focus on the STEM fields.

Data were collected using multiple methods, beginning with the WoS, in order to obtain the citation history and full-text content of articles. Specifically, the DCI of the WoS served as a starting point for gathering records of cited research data providing a single access point to over 350 data repositories worldwide that house over 7.4 million records and 6.5 million citations (Clarivate Analytics, 2018). The DCI links datasets and published research articles and tracks the citation of data while also encouraging its bibliographic citation.

This study used the DCI as evidence of data sharing because it tracks published quality research data (i.e., recording of citation history) across multiple disciplines around the world, thereby allowing easy access to influential data repositories, data sets, data studies, and software. Thanks to these features, I was able to search the DCI directly in order to obtain published quality research data regarding the citation history of data repositories, datasets, data studies, and software worldwide, again from a single access point. In this way it was possible to view and access journal literature, conference proceedings, and books as well as datasets, data studies, and software. A dataset citation includes such components as author, title, year of publication, publisher, edition or version, citation history, and access information (e.g., a URL or other persistent identifier such as a DOI). Within the DCI, I selected higher-level categories (i.e., WoS research areas) rather than lower-level ones (i.e., WoS subject categories) because the DCI's approximately 150 research areas contain more datasets than its 250 or so subject categories.

With regard to data reuse, I obtained the full text of publications (e.g., articles and conference proceedings) online, either directly or through major databases accessible from the University of Wisconsin Milwaukee Libraries website (http://uwm.edu/libraries). When any portions of these publications were unavailable electronically, I obtained print versions from a library, either directly or through inter-library loan. I excluded any documents for which the full text could not be obtained by any of these means. The citation history for each citing article was collected through WoS. All types of documents, including journal articles, conference proceedings, and books, were considered. I used the WoS journal classifications in preference to other classification schemes because "The ISI journal classification system, while it does have its critics, is based on expert judgment and is widely used" (Boyak, Klavans, & Börner, 2005, p. 360). Further descriptions of the documents can be found in the description of the sampling strategy below.

Table 4 summarizes the data collection methods employed in this research at the data, article, discipline and interdisciplinary levels. At the data-level, the information collected was used to study the sharing of published quality research data in the DCI database. Article-level data were reused for the analysis of citing articles. Discipline-level data were used to study both data sharing and reuse. Lastly, interdisciplinary-level data were used to study citation interactions across STEM fields. A detailed description of the research areas at the discipline-level can be found in Table 5.

Table 4 Summary of data collection methods

| data collection | description | collection method |
|---|---|---|
| **data-level** | published data in the DCI (i.e., data sharing) | DCI |

| | | |
|---|---|---|
| **article-level** | published citing articles in the WoS All collections (i.e., data reuse) | DCI, WoS All collections |
| **discipline-level** | the prevalence of data sharing and reuse | Research Areas both in the DCI and the WoS All collections (Table 5) |
| **interdisciplinary-level** | the citation interactions of STEM fields in the WoS All collections | DCI, WoS All collections |

## 3.3. Sampling Strategy

The research population ranged from citing articles in the WoS to highly cited datasets in the DCI, the focus was on STEM fields where research data are shared and reused.  In this study, all of the records were collected from eight different disciplines representing STEM fields; these disciplines thus define the scope within which the findings regarding scholarly communication are reported.

In order to identify the disciplines to be studied within the STEM fields, I compared the major NSF discipline codes (National Science Foundation, 2010), the Research Areas of the WoS All Collections (Clarivate Analytics, 2012), and the research areas of the DCI (Clarivate Analytics, 2016), as can be seen in Table 5. Based on the comparisons (also in Table 5), this study used those WoS research areas for sampling. The eight disciplines were astronomy/physics, biological sciences, chemistry, computing, earth sciences, engineering, mathematical sciences, and technology. Technology was included despite the fact that the NSF major discipline code does not include it as a discipline because STEM by definition includes it. Further, astronomy and physics were merged although the NSF major discipline code separates them because they are combined in the same college/department in many universities. Regarding WoS Research Areas, interdisciplinary areas were not included because their breadth made it difficult to assign them to any one of the identified disciplines. Further, some research areas were not included owing to the selected cut-off point. At a given cut-off point, the total number of records in the DCI for each research area decreased from 4,000 records to 2,000 records (i.e., there were only 2,000 total records or fewer in the DCI for each of these research areas).

Table 5 Comparisons between the NSF major discipline and research areas of the WoS (both WoS All Collections and the DCI)

<table>
<tr><th>NSF<br>- major discipline</th><th>WoS All Collections</th><th>DCI</th></tr>
<tr><td>astronomy</td><td colspan="2" rowspan="2">Astronomy &amp; Astrophysics, Physics, Spectroscopy</td></tr>
<tr><td>physics</td></tr>
<tr><td>biological sciences</td><td colspan="2">Genetics and Heredity, Biochemistry &amp; Molecular Biology, Biotechnology &amp; Applied Microbiology, Cell Biology, Developmental Biology, Evolutionary Biology, Marine &amp; Freshwater Biology, Mathematical &amp; Computational Biology, Microbiology, Plant Sciences, Reproductive Biology, Environmental Sciences &amp; Ecology, Biodiversity &amp; Conservation, Research &amp; Experimental Medicine</td></tr>
<tr><td>chemistry</td><td colspan="2">Chemistry, Crystallography</td></tr>
</table>

| | |
|---|---|
| computing | Computer Science |
| earth sciences | Geology, Oceanography, Geochemistry & Geophysics, Meteorology & Atmospheric Sciences, Water Resources |
| engineering | Engineering |
| mathematical sciences | Mathematics |
| - | Technology |

I used the DCI database to identify the authors (individuals) who have been most active in publishing their data in the DCI. To be more specific, the 30 most productive authors of published documents (datasets, data studies, repositories and software) in each research area were selected. The same process was conducted across diverse research areas. There is no general agreement regarding the appropriate sample size, that is, the appropriate number of groups and of members within each group suitable for multilevel analysis (Raudenbush & Bryk, 2002).

In order to identify citers, I used citing articles to highly cited datasets in the DCI. Influential authors were identified as the first authors of the most highly cited published documents (e.g., datasets, data study, software and repository) in the DCI. The first author was assumed to be the one who made the most significant contribution and the last author was the senior researcher with the most prestigious reputation (Wren et al., 2007). Wren noted that the last author may be the senior researcher with the most prestigious reputation, but this is not always the case, so the first author was selected. For cases in which there was more than one highly cited dataset by the same first author, then the next dataset on the list was selected.

The identification of citers was important because citer-based measures can provide complementary means to citation-based measures to assess higher levels, such as the institution or research group (Ajiferuke, Lu, & Wolfram, 2010). The disciplines of this study (i.e., STEM) were ones in which researchers in the same institutional and research groups can exert an influence through hyperauthorship that extends across multiple disciplines (i.e., is interdisciplinarity) rather than being discipline-specific for big science. As Ajiferuke, Lu and Wolfram have found, there are significant differences between citer- and citation-based results, and "citation measures may not adequately address the influence, or reach, of an author because citations usually do not address the origin of the citation beyond self-citation" (Ajiferuke, Lu, & Wolfram, 2010, p. 2086). These differences were given careful consideration in answering RQ3.

## 3.4. Data Analysis

An exploratory approach was appropriate for this research to understand and answer the research questions of this study because the phenomenon of research data sharing and reuse on data citation was a relatively new area and is in its infancy. Due to the relative reflection of data sharing and reuse on data citation, which itself was not as well documented as formal bibliographic citations, semi-automatic examination of content analysis was appropriate for this exploratory research. As Thelwall (2014, p. 65) noted, "Content analysis involves manually assorting a sample of comments into researcher-defined categories. It is most suitable for exploratory investigations into new phenomena or context".

Employing the methods used by Park, You and Wolfram (2018), a semi-automated method using text searching was applied in order to identify candidate examples of data reuse in publications. Automatic detection of terms/phrases associated with data reuse (Table 10 ) was used by manual verification. Strictly manual methods to identify candidate occurrences of data reuse would be labor intensive for the corpus of publications to be analyzed. Once candidate instances of data reuse had been identified automatically, I manually examined each identified article for evidence of formal (i.e., cited) and informal (i.e., mentioned in passing or implied) data reuse and sharing, whether in the references, main text, acknowledgements, supplementary information, or author information section.

As shown in Table 11, in order to verify the reliability of my data analysis during the content analysis, I used another PhD degree holder in social sciences to assess inter-coder agreement. Since it was impractical and far more time-consuming to repeat the full coding of the citing articles that I judged, (i.e., over 15,000 instances of data sharing and reuse from published articles), 10% of the citing articles were assessed by the second coder. When conducting content analysis, the validity of the human judgments was an important issue to make the identification of data sharing and reuse. In order to ensure the validity of the human judgments, an expert who possesses an understanding of the scientific articles in an academic context (i.e., PhD degree holder) was used. This step helped me establish a level of consistency throughout the research.

### 3.4.1. Data sharing (RQ1: How prevalent is data sharing in different disciplines as measured by formal data citation in STEM fields?)

In answering RQ1, regarding the prevalence of data sharing in various STEM disciplines as measured by formal data citation, descriptive data analysis served as a means to examine data sharing practices across multiple disciplines, each of which has its own data sharing practices (Helbig, Hausstein, & Toepfer, 2015; Torres-Salinas, Jiménez-Contreras, & Robinson-García, 2014) and is therefore deserving of separate study. The total numbers for the shared research data in a range of research areas in the DCI were displayed in a graphic format. I used the "advanced search" feature of the WoS and then such Booleans as "AND" and "OR" to limit each STEM discipline to several research areas. I then recorded the number of times each piece of research data was shared in the DCI. The same procedures were conducted for all of the selected STEM disciplines.

### 3.4.2. Data types (RQ2: What types of STEM research data are formally cited most often?)

In order to answer RQ2, I applied descriptive analysis to examine what types of STEM research data were more often cited. To be specific, I downloaded 100,000 records from the DCI. As mentioned above, 100,000 records were the maximum number that the DCI allows users to download per discipline, which was the same feature as WoS. The cut-off year was 2003 since 2003 was the earliest year when data sharing was required by a federal funding agency in the United States (e.g., NIH since 2003, NSF since 2011) or launched (in 2013 in the United States, 2015 in the United Kingdom, and 2016 in the European Commission). The same procedures were conducted for all STEM disciplines and then saved in a tabular form. Only records with more than one citation count were sorted to identify data types that was most highly cited. I then used Microsoft Excel by using the Pivot Table feature. In each discipline, the top 10 most highly cited data types were identified.

### 3.4.3. Author self-citation and recitation (RQ3: How do author self-citation/recitation practices differ across STEM disciplines?)

To answer RQ3, I applied citer-based analysis as discussed previously. Citer-based analysis was appropriate for examining the various manifestations of self-citation, including author self-citation and recitation. This aspect of the study was important because, while most research data citations continue to be of the self-cited variety (Park & Wolfram, 2017; Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2016), bibliographic reference analysis usually did not address the phenomenon (i.e., the origin of a citation of co-authors' work or recitation in the data citation environment). Moreover, generally speaking, the influence of a work is directly proportional to the number of people who have cited it.

The definition of self-citation has been extended "to include citations originating from publications authored by one of the coauthors of the cited publication of interest, or coauthor self-citations" (Ajiferuke, Lu, & Wolfram, 2010, p. 3). Nevertheless, citations usually do not address their origin beyond self-citation. Also, the DCI did not report the number of self-citations. As the scholars just cited noted, citer-based analysis, which is a form of author research impact analysis, provided the number of unique authors (i.e., individuals) who have cited a given author.

In order to address the issues of self-citation, author self-citation and recitation, this study applied citer-based analysis similar to the method used by Lu, Ajiferuke and Wolfram (2014). This approach measures the impact of an author's research, whereas traditional (e.g., bibliographic) citation-based analysis may not measure author self-citation and recitation in data citation environments. Citer analysis involved measuring author impact based on the number of citers, as opposed to the number of citations (Ajiferuke, Lu, & Wolfram, 2010). Park and Wolfram (2017) found, using citer-based analysis, that self-citation, including author self-citation, was prevalent in data citation in genetics and heredity. The unique publications that cited a publication of interest were represented by the number of citations for each publication. Functions for both the DCI and All Collections of the regular WoS databases were used for data analysis, which followed methods used by Ajiferuke, Lu and Wolfram. All publications by each author were identified using an author search of the WoS. The data for each author's publications were then tabulated and stored.

In order to omit self-citations, the "create citation report" feature provided by both the DCI and All Collections of the regular WoS was analyzed by collecting the bibliographic references for the citing articles for each publication. I used the "analyze results" feature provided by the two databases of the regular WoS in order to identify the citers for each publication. All of the retrieved results (i.e., all of the citing articles) of the sampled authors who had been identified in the DCI as the most-cited in each research area were saved in tabular form, and the average citations "with self-citation" and "without self-citation" were analyzed comparatively.

To examine the associations between and across shared research data and the author self-citation or recitation in the 8 STEM fields, a Kruskal-Wallis test was conducted. A one-way ANOVA test was not suitable for comparing the groups because of the violation of the ANOVA assumptions related to unequal standard deviations.

### 3.4.4. Data reuse (RQ4: How do data reuse practices differ across STEM disciplines?)

In answering RQ4, in order to identify articles (i.e., citing articles at the article level) in the WoS database, systematic random sampling was conducted. Systematic random sampling is a random sampling technique where the first item in a list is randomly selected from the first *n* items on the list and every *n*th observation thereafter is selected in the dataset. In this study, systematic random sampling was preferred to simple random sampling because systematic random sampling "ensures a high degree of representativeness" (Gravetter & Forzano, 2012, p. 147). Although simple random sampling removes bias from the selection procedure, "in the short run, however, there are no guarantees" (Gravetter & Forzano, 2012, p. 146). As Gravetter and Forzano stated, if I select 11th observation, the bias is against choosing observation 12th, 13th, and 14th, which is skewed and distorted sample regarding simple random sampling. The advantage was simplicity of implementation; the drawback was failure to account for possible clumping characteristics within a population. In informetrics, it is necessary to consider whether sampling is performed at the item level or at the informetrics source level, since the former may "not provide a complete portion of any single source if straight count sampling is used, whereas sampling at the source level reduces the number of sources included in the study set" (Wolfram, 2003, p. 73).

In order to track citations, I sorted all records by date in the All Collections of the WoS. This study applied 2003 as a cut-off point for the reason cited in Section 3.4.2 above. All of the retrieved results (i.e., all of the citing articles) of the 30 authors in each research area were saved in tabular form and subjected to systematic random sampling of every 10th citing article (e.g., from the 1st, 11th, 21st, 31st, and so on up to the 91st) of the 30 authors. These citing articles were collected from the "All Collections of the WoS". When the citing articles could not be obtained, the next citing article record from the list was selected (e.g., 1st, 11th, and 22nd for a situation in which the 21st article in the series could not be obtained from the WoS). Using both the DCI and "All Collections of the WoS," the citing articles for each publication (constituting the data) were collected by means of the "create citation report" feature provided by the WoS. Next, the "analyze results" function for these citing articles was used to identify the citers for each publication. Disambiguation of the authors' names was based on the output for the citer data produced by the

WoS (Lu, Ajiferuke, & Wolfram, 2014) because only slight differences, of a few percent or less, were found (Smalheiser & Torvik, 2009).

In order to identify and collect instances of data reuse that were embedded within publications that were not formally included as citations, this study applied text searching based on the appearance of the selected words and phrases. This method was precise, allowing identification of only those information resources that was relevant to my information needs by removing resources that may not include data reuse, such as documents (i.e., articles) without indicating terms/phrases. The potential for larger samples being captured using text searching techniques, as opposed to strictly manual searching, ensured that the data citation research would be representative and diverse.

The publications collected from the University of Wisconsin Milwaukee libraries (http://www.uwm.edu/libaries) included substrings of terms/phrases indicating the possibility of data reuse, such as "acquaintances," "donated from/by," and "repositories." Thus, for instance, the indicating term/phrase "repositories" along with its substrings (e.g., "repository," "repository numbers") were searched and collected automatically. Table 6 lists the terms/phrases used for the search strings. I analyzed another research area, physics, in order to confirm whether currently identified terms/phrases can be applied in other research areas. Terms/phrases indicating data reuse and sharing in the field of genetics and heredity (Park & Wolfram, 2017) were listed in Table 6. Except for "NIH" (National Institutes of Health), terms can be regarded as generally applicable to other research areas. Another seven disciplines (Table 5) were analyzed to identify indicating terms/phrases that could be applied other disciplines. Five sample documents published in these seven disciplines were analyzed in order to identify terms/phrases indicating data reuse in these fields (Section 4.1).

Table 6 Terms/phrases indicating data reuse and sharing in the field of genetics and heredity (Park, & Wolfram, 2017)

| **indicating terms/phrases of data reuse** | **indicating terms/phrases of data sharing** |
|---|---|
| “commercial,” “Corp.,” “database,” “donated from/by,” “gift,” “.gov,” “Inc.,” “indebted,” “lab/laboratories,” “Ltd.,” “obtained from,” “purchased from,” “repository,” “repository numbers,” “samples,” “sample sets,” “survey” | “accession #,” “deposited,” “National Institutes of Health,” “NIH,” “project website,” “publicly available,” “repository,” “stored,” “suppl,” “supplemental,” “supplemental material” |

Table 7 displays the summary of articles associated with research data in each discipline and their total instances of citation, thereby capturing the phenomenon of data citation based on the modified terms/phrases that indicate data sharing and reuse. This step was conducted based on the findings from a pilot study (Section 4.1). In STEM, the total instances extracted by using indicating terms for data sharing and reuse included 15,263 unique instances from 705 articles. Total numbers of citing articles and total instances varied depending on disciplines. This step was conducted to examine the phenomenon of research data and their associated scientific publications. The instance disparities in each discipline in terms of data sharing and reuse may affect the analysis. When looking at the number of associated articles versus their associated research data sharing and reuse, the large number of instances skews the results. For instance, the skewed instances in biological sciences (44.35% of all STEM) affect the result. This could mean that more researchers in biological sciences were sharing and reusing for crediting data sharers, or it could mean that the policies of publishers and funding agencies for data sharing were stricter in biological sciences.

Data sharing is seldom inspired by the data sharing mandates of funders (Couture, Blake, McDonald, & Ward, 2018). The NIH's relatively early data sharing mandates which date back to 2003 (compared to 2011 for NSF), more instances in biological sciences can be explained. Yet this may not be true, journals in biological sciences mandate more to submit research data to repositories than other disciplines can be another reason. Disciplinary differences for journal policies, repositories and normative pressure had significant positive effect on data sharing in scientific disciplines (Kim & Stanton, 2015). Strong journal policies for data sharing are associated with increased data sharing for all first and last authors for high-impact journals in biomedical microarray studies (Piwowar & Chapman, 2010). The genomics community mandates depositing dataset in repositories (Costa, Qin, & Bratt, 2016). Adopting different pace of data sharing and reuse practices can be another reason. The early expansion of these practices observed in the genomics and astronomy communities (Borgman, 2012) is paralleled, in the biological sciences as well as in astronomy/physics.

Table 7 Total associated articles and their total instances

| **discipline** | **total numbers of associated articles** | **total instances** |
|---|---|---|
| astronomy/physics | 78 | 1,935 |
| biological sciences | 235 | 6,768 |
| chemistry | 119 | 1,083 |
| computing | 14 | 499 |
| earth sciences | 121 | 1,796 |

| | | |
|---|---|---|
| engineering | 56 | 885 |
| mathematical sciences | 38 | 1,153 |
| technology | 44 | 1,142 |
| **grand total** | **705** | **15,261** |

### 3.4.5. Interdisciplinarity (RQ5: To what extent do STEM disciplines support interdisciplinary data citation?)

Answering RQ5, regarding the extent to which the various STEM disciplines support interdisciplinary data citation, served as a general glimpse of the interdisciplinary impact of citations. Answering it involved analyzing interdisciplinary interactions among diverse disciplines in the time since the advent of open science. Citation has been used to monitor the evolution of interdisciplinarity because citation networks at the level of published articles across disciplines reflect the flow of knowledge. Interdisciplinary knowledge is transferred through cross citations as well as papers that appear frequently in diverse disciplines.

In this study, I used the term ‘interdisciplinary’ rather than multidisciplinary or transdisciplinary. The reason is that the term interdisciplinary is widely and ambiguously used for research across various areas such as scholarly communications, industrial sectors and technological fields although the terms inter-, multi- and trans-disciplinary is between, beyond or across disciplines (Rafols & Meyer, 2010). The terms, multidisciplinary, interdisciplinary and transdisciplinary, have been initially introduced by Thomlinson (1983). ‘Multidisciplinary’ which indicates works with several disciplines (Whitfield & Reid, 2004), is a process that provides a juxtaposition of disciplines as additive not integrative (Klein, 1990). ‘Interdisciplinary’ works between several disciplines (Whitfield & Reid, 2004) and builds a new level of discourse and integrates knowledge (Klein, 1990). ‘Transdisciplinary’ works beyond (Nicolescu, 1998) and across (Whitfield & Reid, 2004) several disciplines and examines the dynamics of whole systems and is a holistic scheme of subordinate disciplines (Klein, 1990). Aboelela, et al. (2007) further compared multidisciplinary, transdisciplinary and interdisciplinary. A multidisciplinary team includes researchers from two or more disciplines and work on the same questions without much interaction although separate publications by researchers from each discipline may be produced. Transdisciplinary includes two or more distinct academic fields with shared publications to solve complex problems, probably using some new languages developed to translate across traditional lines. Interdisciplinary includes two or more distinct academic fields, with shared publications by using language intelligible to all involved fields.

In order to measure the interdisciplinarity of data citation received for each discipline, the citation of a paper in each field (i.e., citing articles with at least one citation) had been used. This study applied journal citations for the following reasons. Citation analysis was widely used for measuring the interdisciplinarity of research output since citation data can reveal to us past, present and future activities in science (Garfield, Malin, & Small, 1978). Citation analysis allowed how one research field borrowed the knowledge from another field. Journal analysis reveals the integration of different research fields because those fields share publication outlets (i.e., journals). The hypothesis was that being cited with multiple fields can be an evidence of the interdisciplinary nature of publications than those being cited with single field.

Clarivate Analytics (2018) assigns the 11,700 journals that it indexes to one or more Subject Areas, Research Areas, and Essential Science Indicators (ESIs). These assignments can be applied to the journals in which the citing articles appear, thereby providing an indication of the interdisciplinary impact of citable datasets. The 2018 ESI categorizations of citing articles for the journals were used instead of the Subject Areas and Research Areas because Clarivate Analytics assigns each journal in the ESI database to one, and only one, of 22 ESI research fields, thereby avoiding ambiguity. The smaller number of categories also allowed for a more manageable subject categorization of citing articles. The ESI data for the citing articles and their associated journals were entered into a relational database management system, Microsoft Access, for matching purposes. The relatively recent time-frame of the study—again, beginning in 2003—for publications indexed by scholarly databases (i.e., the WoS) can itself mitigate challenges associated with studying a scholarly database because significant time may pass from the publication of an article to the appearance of the data referenced therein in a scholarly database (Bollen et al., 2009). Another concern was the continuous updating of subject classifications of journals as a means to “overcome the birth of new journals and to identify the emergence of new disciplines” (Gómez, Bordons, Fernández, & Méndez, 1996, p. 227), for disciplines and journals alike have been subject to rapid change. Use of the 2018 ESI categorization, the most recent version available, however, can mitigate this concern as well.

The assumption was that the ESI research disciplines were well organized and did not introduce any significant changes into the overall analysis. I assumed that research data citation or bibliographic citation indicates scholarly influence or knowledge transfer in scholarly communication. Under these circumstances, richer information regarding authors’ publishing articles in the journals of other fields can be obtained for interdisciplinary disciplines by examining (1) individual field of a journal and (2) field of journals that frequently cite it. Thus, for instance, a journal in the subject category of physics contributes to astronomy/physics and is likely to be cited by journals classified the subject assignments physics and astronomy. In other words, publications in a given field can be considered indicators of the diffusion of particular discipline.

Three aspects of diversity in interdisciplinarity have been identified, namely variety, balance, and disparity (Leinster & Cobbold, 2012; Stirling, 2007; Zhang, Rousseau & Glänzel, 2016). Variety refers to the number of disciplines to which references made in a paper can be assigned, balance to the evenness of the distribution of the discipline classification, and disparity is measured as the distance between the disciplines to which the references are assigned.

In order to measure interdisciplinary data citation using a single formula, I applied Leydesdorff's (2018) interdisciplinarity calculation along with the Gini-index and the number of ESI categories represented in the citing articles for each discipline. This combination made it possible to measure diversity while distinguishing variety, balance, and disparity. The Gini-index alone was insufficient because, while it indicates balance (Nijssen, Rousseau, & van Hecke, 1998), it does not indicate variety (Leydesdorff, 2018).

The following formula measures the three aspects of diversity such as variety, balance and disparity (Leydesdorff , 2018). The raw matrix was used to create a relative frequency asymmetric matrix for each STEM discipline and ESI field.

$$Div_c = \frac{n_c}{N} * Gini_c * \left[\frac{\sum_{i=1,j=1,i\neq u}^{i=n_c,j=n_c} d_{ij}}{\{(n_c\,(n_c-1))\}}\right]$$

The three parts are used to calculate Leydesdorff's formula.

In the first of the three parts, the number of represented fields was divided by the number of 2018 ESI fields, which was 22. In the second, the Gini-index was calculated using the simplified formula to measure the inequality of the distribution as an indicator of balance. The relative frequency was $\frac{n_c}{N}$ . The variables *i* and *j* represented each observation in the cells along the vector. These *i* and *j* permutations of the cells excluded the main diagonal. To compute the Gini coefficient, I used Leydesdorff's (2018) simplified calculation where $n$ was the number of elements observed, *i* was the rank of values in ascending order and $x_i$ was the number of citations of element *i* in the ranking. Each of the three components of the formula varied between 0 and 1. For example, a Gini-index of 0 indicates the citations are equally distributed over the papers and a Gini-index of 1 indicates a single paper receives all citations.

To be specific, for each observation in a given field, ascending order was applied, thus, for instance, the areas that contribute 0 went first and contributed nothing. All of these numbers were then summed up for the numerator total. The denominator was always 22 (i.e., the number of 2018 ESI fields) because the total probabilities always total 1 multiplied by *n*, which is 22.

$$Gini = \frac{\sum_{i=1}^{n}(2i-n=1)x_i}{n\sum_{i=1}^{n}x_i}$$

In the third part, a normalization factor, was applied. In order to warrant the disparity as weightings between 0 and 1, 1 minus the cosine similarity between data elements was used for normalization.

$$\text{Similarity} = \cos(\Theta) = \frac{\mathrm{A}\cdot\mathrm{B}}{\|A\|\|\mathrm{B}\|} = \frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^{\,2}}\sqrt{\sum_{i=1}^{n}B_i^{\,2}}}$$

A cosine calculator (https://calculator.vhex.net/calculator/distance/cosine-distance) was employed to compare vectors. Each vector consisted of the relative frequencies of the 22 ESI categories in the citing articles for each discipline. A pairwise comparison of the relative frequencies across the non-zero ESI categories resulted in the creation of an $n_c$ x $n_c$ matrix. The $n_c$ was the non-zero ESI categories. The resulting value (i.e., 1 minus the cosine value) was placed in the corresponding cell for each pair. As a symmetric matrix, each pair was compared once and the corresponding value was entered in the mirror cell. The resulting matrix (i.e., $n_c$ x $n_c$) was used for the normalization factor calculation. To be specific, the values for each cell calculation were summed up by taking the distance value in the cell and dividing it by $n_c$ x $(n_c - 1)$. Thus, for instance, the denominator for the astronomy/physics discipline was 42 (i.e., 7 x 6) since the value 7 was $n_c$ and 6 was ($n_c$ -1).

Finally, the resulting diversity value for each of the fields was multiplied.

## 3.5. Summary of the Research Design

### 3.5.1. Data Analysis Strategies

Table 8 summarizes the research questions, data collections and the data analysis strategies used. Again, the purpose of this study is to improve the associations among the data, article, discipline and interdisciplinarity-levels of research data citation. Multiple data analysis techniques were employed in order to answer the research questions.

Table 8 Research questions and data analysis

| **category &<br>research question** | **data collection** | **data analysis** |
|---|---|---|
| **data sharing**<br>**RQ1** | Published data in the DCI | descriptive analysis |

| | | |
|---|---|---|
| **data type**<br>**RQ2** | Published data in the DCI | descriptive analysis |
| **self-citation**<br>**RQ3** | Published citing articles in the WoS All Collections | citer-based analysis,<br>Kruskal-Wallis |
| **data reuse**<br>**RQ4** | Published articles in the WoS All Collections | content analysis |
| **interdisciplinarity**<br>**RQ5** | Published articles in the WoS All Collections | Gini-index,<br>Leydesdoff's interdisciplinarity calculation |

### 3.5.2. Validity

There is rich body of literature discussing the validity of using citations to measure research impact, and "The standard test of the validity of evaluative citation counting is comparison with peer evaluation, including the evaluations made in awarding of prizes and grants" (Lercher, 2013, p. 455). Correlations between citation counts and other measurements of influence, such as peer reviews and rewards, have thus been actively studied. Clark (1954), for example, found that citation counts correlate strongly with the assessments of the most influential researchers in the field of psychology.

Validity may be compromised when the peer-review process for evaluating both quality research datasets (e.g., those in the DCI database) and articles (e.g., those in the regular WoS database) is limited to subsets of the entire research output. The validity of a measurement procedure refers to “whether the procedure actually measures the variable that it claims to measure and threats to validity include “any component of a research study that introduces questions or raises doubts about the quality of the research process or the accuracy of the research results” (Gravetter & Forzano, 2012, p. 167). The validity of the various journal similarity measures and the corresponding maps is generally approached using the WoS journal classifications for the validation of science maps because “The ISI journal classification system, while it does have its critics, is based on expert judgment and is widely used” (Boyak, Klavans, & Börner, 2005, p. 360).

### External Validity

External validity is related to generalizability; it refers to “the extent to which we can generalize the results of a research study to people, settings, times, measures, and characteristics other than those used in that study” (Gravetter & Forzano, 2012, p. 168). In this study, the research data and articles were collected from the STEM research areas, including science, technology, and engineering and mathematics, which represent different disciplines. Therefore, the outcome of this research can reveal these disciplines in scholarly communication. Also, the datasets used in informetrics, the datasets used are assumed to represent either a random sample of the overall population under study or the population itself (Wolfram, 2003).

### Internal Validity

Any factors that allow for alternative explanations for the results as a study proceeds represent a threat to its internal validity, which relates to factors that raise questions or doubts regarding the interpretation of the results (Gravetter & Forzano, 2012, p. 170) and is thus a quantity related to the logic and coherency of cause-and-effect explanations. As a measure of internal consistency, this research used Cohen's (1960) kappa coefficient to ensure inter-rater agreement for qualitative items because *k* considers the possibility of the agreement occurred by chance.

## 3.6. Strengths and Limitations

### Strengths

In regard to the strengths of the proposed approaches, to begin with, citer-based analysis may overcome the limitations of more traditional co-citation approaches, including the subjectivity of citers that is inherent in citation-based data (Lu & Wolfram, 2012). A second strength is that prolific authors and co-authors, as well as the coupling frequency, can be identified. Third, the mapping of scholarly communication allowed for the visual interpretation of complex interconnections based on citations and links. The application of informetrics may help in the development of scientific indicators and in evaluating the impact of the scholarly communication process and interdisciplinary relationships. Fourth, informetrics and NLP attributes facilitated the examination of a very large set of research and attributes in the context of data reuse. Finally, the trust and reliability of research data quality was an important consideration for data reusers. Considering this study analyzed quality research data from the DCI, the outcomes of this study may confirm other findings meaningful for data reusers.

## Limitations

Works in which data were associated with journals and indeed, over 90 percent of works with associated data were journal articles (Park & Wolfram, 2017) may hinder inquiries into phenomena in rapidly advancing areas, such as in the hard sciences or computer engineering. In such areas, in contrast with the situation in the humanities or social sciences, conference proceedings are considered of greater importance than journal articles or books. This was in large part because the review process for articles or books may take more than a year, depending on the journal or publisher, which in turn may be due to the policies of high-profile journals that include strict data sharing requirements, while conference proceedings do not currently have strict data sharing policies, though the same is true of books. As Callaghan and colleges (2012) noted, further research should be conducted regarding the scientific validity of the datasets because those datasets cannot be claimed as equivalent as an already established peer-review process for traditional academic publishing (i.e., the scientific quality of the datasets), because creating a mechanism for the full peer review of the scientific publication of datasets are still in the early stages.

This study focused on data citation characteristics found in papers that cite prolific authors in the identified fields. Citations to less influential authors, whose work presumably receives fewer citations, were not investigated.

This study also may have underestimated the total amount of data sharing in all forms because it did not include laboratory or personal websites or direct sharing, such as between personal acquaintances (i.e., peer-to-peer data sharing) or within a collaboration network. Furthermore, the reliance on indicator terms to identify potential examples of informal and formal data citation may not reveal all occurrences of data citation.

## 3.7. Summary

The data collection methods and research design outlined in this chapter formed the basis for a wide-ranging investigation of data sharing, reuse and citation phenomena in STEM disciplines, as documented by the DCI.

# Chapter 4 RESULTS

This chapter outlines the findings of the study. Informetric methods and text searching provided useful analytical tools for exploring data citation. This mixed method study was approached by combining quantitative approaches used in informetrics and qualitative semi-automatic content assessment. One contribution of this dissertation, thus, was to establish a methodological framework. Specifically, a refined research model was developed with reference to key previous works on data citation, data sharing, and data reuse, in particular those that identify groups of factors. The data analysis methods to be used were primarily quantitative; however, there is a qualitative component in the evaluation of data reuse.

## 4.1. Pilot Study

I conducted a pilot study of eight STEM in 8 disciplines (Table 5) because the terms currently identified were drawn from previous studies that focused on the biomedical fields (Park & Wolfram, 2017). This step was conducted in order to identify any missing or new terms not included in previous research and to generalize various terms and phrases. For each of the eight disciplines, 5 articles were examined, totaling for a total of 40 published articles. in STEM (8 disciplines × 5 articles each).

Table 9 Identifying new indicating terms of data sharing and reuse in STEM fields.

| disciplines | examples | previously identified terms | newly identified terms |
|---|---|---|---|
| astronomy & physics | - This research has **benefited** from the SpeX Prism Spectral Libraries, maintained by Adam Burgasser at **http://**www.browndwarfs**.org**/spexprism. | benefited, http:// | benefited, http://, .org |
| biological sciences | - **Data availability**. The **data sets** generated during the current study are **available** at the **database** of Genotypes and Phenotypes (dbGaP) under **accession** phs001273.v1.p1. | accession, available, data availability, database, data sets | availability, data, data sets |
| chemistry | - Supporting Information **Available**: Optimized coordinates and theoretical IR spectra. This material is **available** free of<br>- charge via the Internet at **http://**pubs.acs.**org**. | available, http:// | .org |
| computing | - PI-PLC was **purchased from** Sigma<br>- Chromosome coordinates for most of the human genome elements analyzed here were **obtained from** the UCSC Table **Browser** | browser, http://, obtained from, | browser, .ed u |

| | | | |
|---|---|---|---|
| | (**http:**//genome.ucsc**.edu**/cgi-bin/hgTables?hgsid=357122457). | purchased, purchased from | |
| earth sciences | • At the NEEM site both 10Be and NO3 − **data** are **available**, although derived from two separate neighboring cores with relative age uncertainties of the order of a few years.<br>• Sea-ice concentration **data** are the NASA Bootstrap SMMR-SSM/I combined **dataset** from the US National Snow and Ice **Data Centre** {**http:**//nsidc**.org**; Comiso, 1999}.<br>• Long-term atmospheric CO2 concentrations were **provided by** the National Oceanic and Atmospheric Administration’s (**NOAA**) Earth System Research Laboratory (**http:**//www.esrl.noaa**.gov**) | data, available, data center, http://, NOAA, provided by | data, data center, .gov, NOAA, .org, provided by |
| Engineering | • Five adult Eigenmannia virescens (length 12–15 cm) were **obtained from** a **commercial** vendor and housed according to published guidelines [28]. | commercial, obtained from | - |
| mathematical sciences | • The counties are taken from an ESRI Shape file **downloaded** from the US Census. | downloaded | - |

| | | | |
|---|---|---|---|
| technology | • **Data Availability** Statement: All relevant **data** are **available** on Open Science Framework: **https://**osf.io/cs9c2/. The **data** from the experiment and modeling is part of an R-package, which can be **accessed** here: **https://**cran.r-project**.org**/package=AcousticNDLCodeR. The corpus GECO 1.0 used in this study is **available** from the IMS UniversitaÈt Stuttgart: **http://**www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GECO.en.html | access, available, data, data availability, http://, https:// | data, https://, .org |

**Modifications of terms and phrases**

Table 10 displays 39 modified terms and phrases indicating potential data sharing and reuse that were found through the addition of more STEM disciplines. I used these additional terms and phrases to identify potential data sharing and reuse efficiently measuring data citation. The pilot study, then, confirmed that the key terms and phrases indicating data sharing and reuse for data citation were diverse across disciplines, and it accordingly informed the design of the main study with regard to the assessment of differences in research data practices across and within disciplines.

Table 10 Modified terms/phrases indicating data sharing and reuse for data citation in STEM

| | | |
|---|---|---|
| .com | data sets | obtained from |
| .edu | database | project website |
| .gov | dataset | provided by |
| .org | deposited | publicly available |
| accession | donated by | purchased from |
| acquire | donated from | repository |
| available | downloaded | repository numbers |
| benefited | ftp:// | samples |
| browser | gift | stored |
| commercial | Inc. | Suppl |
| Corp. | National Institutes of Health | Supplemental |
| data availability | NIH | supplemental material |
| data center | NOAA | survey |

I employed a text-searching technique because of the labor-intensive nature and smaller scale of manual methods. In order to detect data citation for data sharing and reuse, I used the 39 modified terms and phrases as described in Table 11 derived from the full text STEM documents.

Prior to the actual data analysis, data cleaning was necessary to avoid any problems and ensure the validity of the data. In this process, outliers—unusual values for variables with the potential to distort the statistics (Tabachnick & Fidell, 2000)—were identified and excluded. Thus, for example, I removed an author with the last name of Lee, as discussed above, because this is a very common name in East Asian countries, making it an outlier in the context of this study. The effect of outliers was in any case marginal because the data analysis involved a large sample.

### Reliability assessment

I created a draft-coding scheme based on my research goals. This step served to reveal valuable patterns that had heretofore gone unnoticed and to avoid introducing personal bias into the identification of data sharing and reuse or into the scale assessment, which could comprise the reliability and validity of the data. Using this draft-coding scheme, as described above, an assistant with a Ph.D. in social science and experience with coding coded 10% of the total instances, which amounted to 1,528 records.

Also, as mentioned earlier, I used Cohen's kappa coefficient to estimate the internal consistency between the two coders (the assistant and myself). As seen in Table 11, I achieved an interrater reliability of 0.814 from the 1,528 records just mentioned. This result indicated sufficient reliability for me to code all 705 of the papers, for a total of 15,261 instances. For the content analysis, I read all of the texts and assigned codes to any that were related to data sharing, reuse, or citation.

Table 11 Reliability test using Cohen's kappa coefficient

**Symmetric Measures**

| | | Value | Asymptotic Standard Error[a] | Approximate T[b] | Approximate Significance |
|---|---|---|---|---|---|
| Measure of Agreement | Kappa | .814 | .017 | 40.268 | .000 |
| N of Valid Cases | | 1528 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

For the content analysis, I read all of the texts and assigned codes to those related to data sharing, reuse and citation. A coding scheme was created to analyze data usage – in which -DS = data sharing, DR = data reuse and /R = repeatedly described for already cited data) and location in the document (A = acknowledgements, AB = abstract, M = main text, R = references and S = supplementary information.

## 4.2. RESULTS

### 4.2.1. RQ1: How prevalent is data sharing in different disciplines as measured by formal data citation in STEM fields?

RQ1 was addressed using population data preserved in the DCI; its objective was as follows.

- Objective: To identify and map various levels of factors that influence data sharing in STEM fields as measured by formal data citation in general.

RQ1 examined disciplinary differences in the STEM fields. Figure 1 reports the prevalence of data sharing in the various disciplines as measured by data citation; specifically, it displays the total records of data sharing in the DCI by year. Data sharing was most prevalent in the biological sciences and least prevalent in engineering. The lowest data sharing values were for 2014, a result attributable to the overall trend in the biological sciences but deserving of further study. Inconsistencies from year to year occurred owing to the unpredictability of the sources of the data. Thus, for instance, the very limited levels of data sharing in 2017 may have been an artifact of the indexing features of the DCI.

I observed, then, distinct disciplinary differences in data sharing across STEM fields, with greater prevalence in the biological sciences and very little activity in astronomy/physics, computing, engineering, and mathematics. Thus, there were more than 3.7 million records of shared research data, such as datasets, software, data studies and, repositories, for the biological sciences, but only around 7,000 for engineering. To be sure, a lower output of datasets does not necessarily mean less data sharing, again owing to differences in data production and use across disciplines (Mongeon, Robinson-García, Jeng, & Costas, 2017). Thus, for instance, a certain discipline may make relatively heavy use of proprietary or sensitive data, such as that gathered from medical patients, that is by nature difficult to share. From this perspective, the growth of the open science movement can complicate data-sharing practices by conflicting with ethical considerations relating to confidentiality.

The published records of data sharing in the DCI, having remained fairly stagnant in the period leading up to 2003, have shown consistent dramatic growth ever since. This result seems to be in line with other patterns in the STEM fields and indicates that research data sharing was, as discussed earlier, not prevalent before major funding agencies began implementing data-sharing policies, the influence of which is thus manifest. (Again, the NIH began requiring a data sharing plan in 2003, and the NSF mandated a management plan for data sharing in 2011, two years later also revising its guidelines to allow biosketches to include references to research data and software).

Figure 1 The prevalence of data sharing as measured by data citation in STEM fields in the DCI

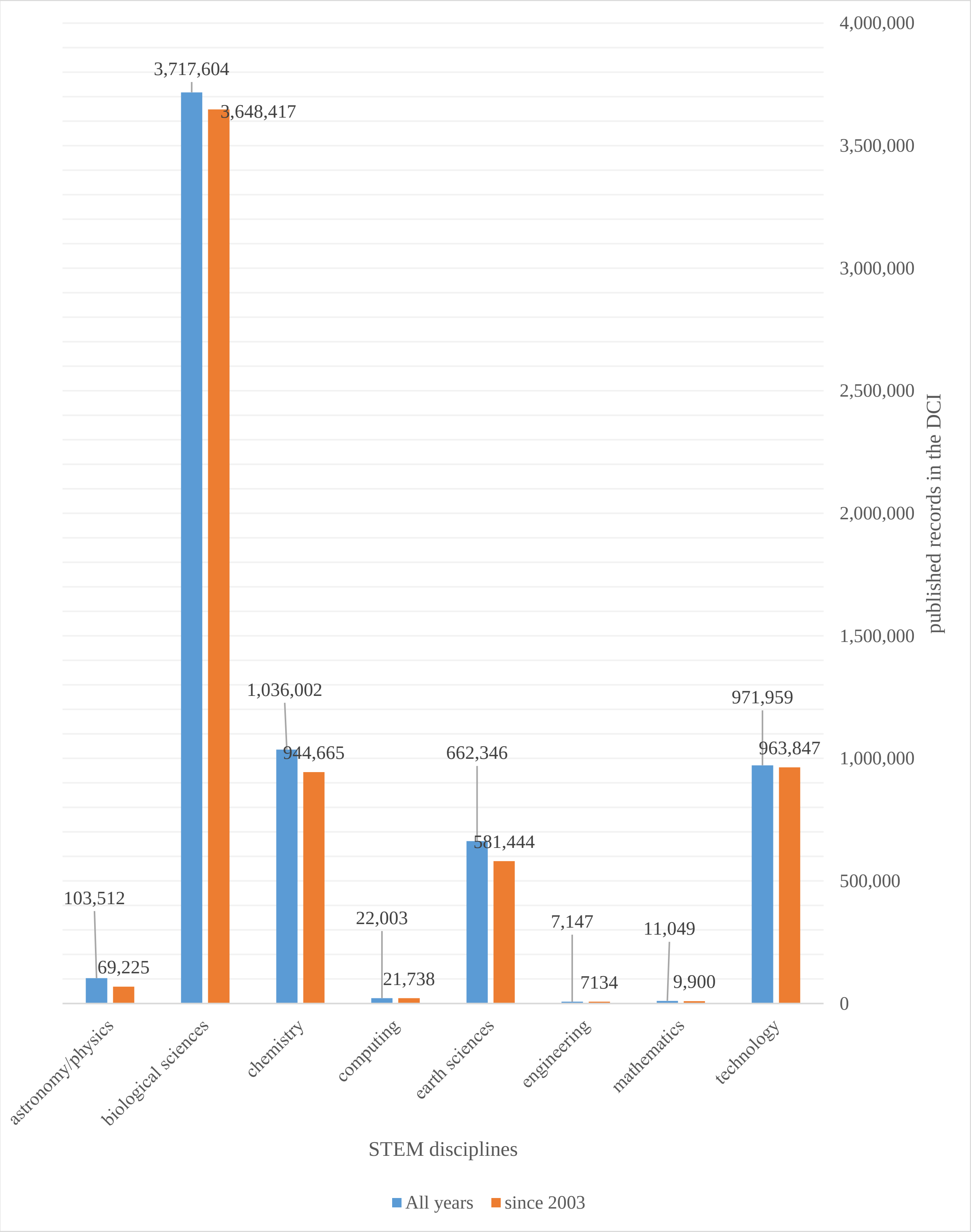


Figure 2 displays data sharing in STEM fields by year as documented by the DCI. An increase is observable, but there was variation from year to year, which is another phenomenon that deserves further study. The rate of data sharing in the earth sciences has decreased steadily since 2008, possibly owing to some of the observed changes indexed by the DCI.

Figure 2 Data sharing in STEM by year in the DCI

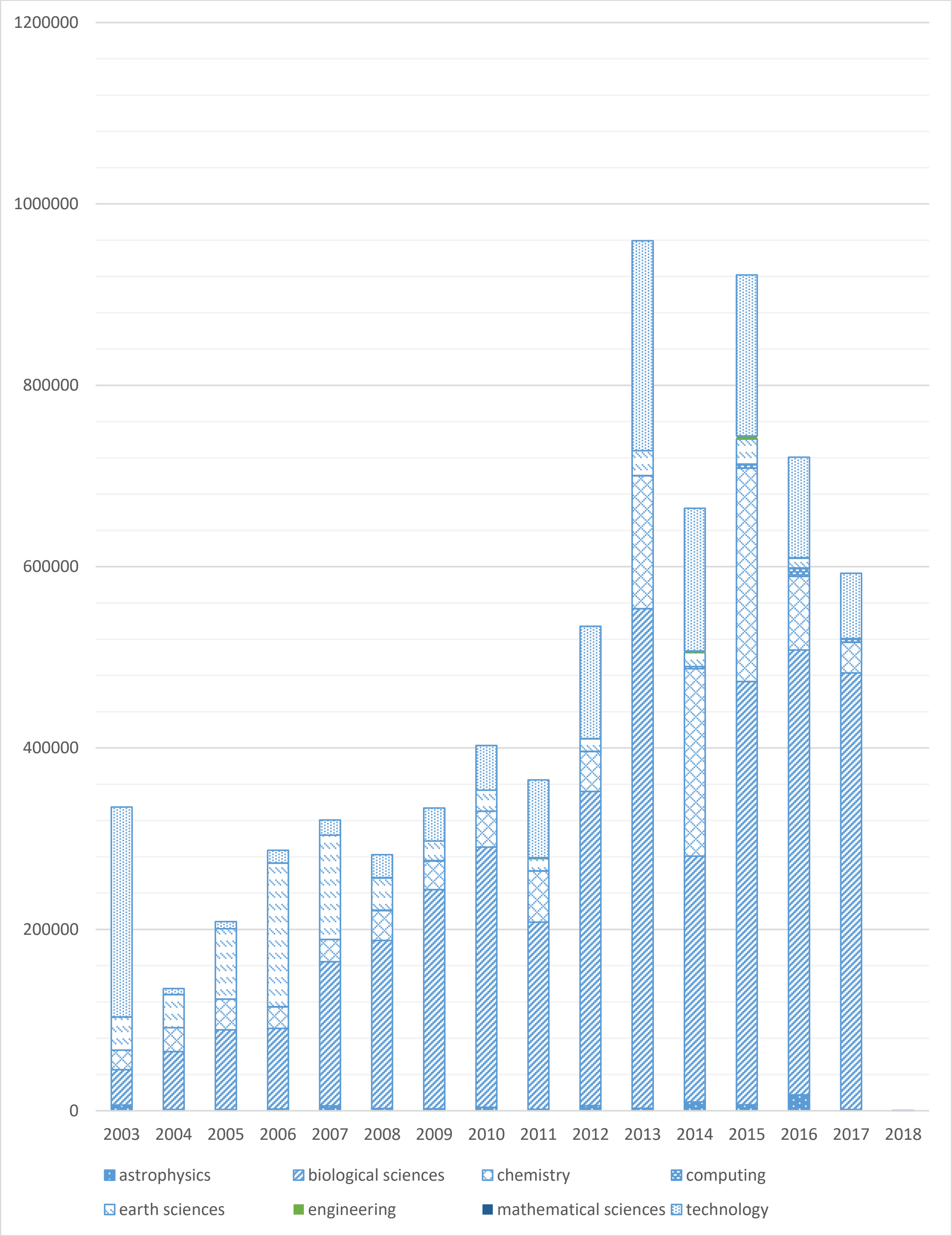


Table 12 presents the results of further examination of the document types in the DCI that are shared among various STEM fields. Swoger has usefully defined a dataset as “a single or coherent set of data or data file provided by the repository, as part of a collection, data study or experiment,” a repository as “a database or collection comprising data studies, and data sets which stores and provides access to the raw data,” and a data study as a “description of studies or experiments held in repositories with the associated data which have been used in the data study” (Swoger, 2012, p. 110).

The distribution of document types also differed by discipline. Datasets were the most commonly shared document type (over 80%) in the DCI except for computing (1.06%), engineering (20.05%), and the mathematical sciences (17.57%). By contrast, datasets were more prevalent in astronomy/physics (86.92%), the biological sciences (87.57%), chemistry (99.15%), the earth sciences (94.78%) and technology (80.67%). In engineering, citations were concentrated in data studies (79.9%), and in computing that were concentrated in software (90.95%).

Table 12 Document types of the STEM fields in the DCI

| **discipline** | **dataset** | **software** | **repository** | **data study** |
|---|---|---|---|---|
| astronomy/ physics | 60,171 (86.92%) | 1,394 (2.01%) | 17 (0.03%) | 7,643 (11.04%) |
| biological sciences | 3,194,748 (87.57%) | 0 (0%) | 70 (0%) | 453,599 (12.43%) |
| chemistry | 936,596 (99.15%) | 0 (0%) | 9 (0%) | 8,060 (0.85%) |

| | | | | |
|---|---|---|---|---|
| computing | 230 (1.06%) | 19,770 (90.95%) | 3 (0.01%) | 1,735 (7.98%) |
| earth sciences | 551,092 (94.78%) | 13 (0%) | 15 (0%) | 30,357 (5.22%) |
| engineering | 1,430 (20.05%) | 0 (0%) | 4 (0.06%) | 5,700 (79.9%) |
| mathematical sciences | 1,739 (17.57%) | 8,155 (82.37%) | 0 (0%) | 6 (0.06%) |
| technology | 762,082 (80.67%) | 10,648 (1.13%) | 37 (0%) | 191,080 (20.23%) |

Table 13 displays the total number of shared research data types for each of the STEM fields in the DCI beyond the four major document types. The biological sciences, earth sciences and technology had the most data types, which engineering and mathematical sciences had the fewest.

Table 13 Total numbers of the shared data types of the STEM fields in the DCI

| **discipline** | **total numbers of the data types in the DCI** |
|---|---|
| astronomy/physics | 86 |
| biological sciences | 100 |
| chemistry | 36 |
| computing | 19 |

| | |
|---|---|
| earth sciences | 100 |
| engineering | 7 |
| mathematical sciences | 6 |

In order to report the results for each discipline as clearly as possible, summaries for the results appearing in Table 14 to Table 21 are divided into two tables of four disciplines each. Table 14 and Table 15 display the top 10 most highly shared/published data types for each discipline in the DCI. Certain types were more widely shared in most STEM fields. In computing and the mathematical sciences, software represented the most frequently shared data type, constituting 91.4% and 82.9% of total records, respectively. Test data represents the most frequently shared data type in engineering by far (99.7%). In contrast, astrophysics did not have a single data type that is most shared. For instance, mass spectral data, the most frequently shared data type represented only 7.2% of the shared data records. Considering a few formats dominant data formats in astronomy (Greenfield, Droettboom, & Bray, 2015), such as Flexible Image Transport System (FITS) files in astronomy (Grosbol, 1988), the findings of only 190 shared FITS files (0.43%) in astronomy/physics and of no dominant data type are remarkable.

Table 14 Data types: The top 10 most highly shared/published data types of the astronomy/physics, biological sciences, chemistry and computing fields in the DCI

| **astronomy/physics** | | **biological sciences** | | **chemistry** | | **computing** | |
|---|---|---|---|---|---|---|---|
| **data type** | **total records** | **data type** | **total records** | **data type** | **total records** | **data type** | **total records** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| mass spectral data | 31,072 | RNA[1] | 931,673 | crystal structure | 754,913 | software | 18246 |
| NMR [2]results | 6,157 | protein sequence data | 525,973 | crystallographic data | 490,252 | code | 1,278 |
| spectral data | 3,723 | SRA[3] | 277,920 | molecular structure | 91,870 | model | 416 |
| software | 1,396 | genomic | 163,349 | crystallographic information | 84,687 | dataset | 3 |
| image file | 233 | images | 113,107 | bacterial carbohydrate structure | 4,298 | raw experimental data | 2 |

[1] Ribonucleic acid

[2] Nuclear magnetic resonance

[3] Sequence Read Archive

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| FITS file | 190 | nucleotide sequencing information | 109,135 | spectral data | 3,720 | other | 2 |
| final output pics | 163 | molecular structure | 75,899 | crystallographic structure | 3,008 | database | 2 |
| data | 107 | processed | 72,717 | dataset | 2,410 | survey and census data | 1 |
| dataset | 63 | FGEM | 72,717 | molecular data | 954 (0.1%) | spreadsheet | 1 (0%) |
| hrcrop | 60 | plant trascription factors and their annotation | 65,536 | molecule | 647 (0.1%) | simulation on MATLAB code | 1 (0%) |
| **totals** | **43,164** | **totals** | **2,408,026** | **totals** | **1,436,759** | **totals** | **19,952** |

Table 15 Data types: The top 10 most highly shared/published data types of the earth sciences, engineering, mathematical sciences and technology fields in the DCI

| **earth sciences** | | **engineering** | | **mathematical sciences** | | **technology** | |
|---|---|---|---|---|---|---|---|
| **data type** | **total records** | **data type** | **total records** | **data type** | **total records** | **data type** | **total records** |
| dataset | 32,975 | test data | 3,749 | software | 8,155 | dataset | 137,375 |
| interactive resource | 22,264 | QCM data[4] | 1 | matrix | 1,640 | fileset | 33,304 |
| GPS dataset[5] | 13,080 | microscopy images | 1 | geoid undulation given on a grid | 35 | image TIFF | 14,558 |
| geoscientific information | 9,108 | GIS vector data | 2 | dataset | 1 | image | 12,591 |

[4] Quartz Crystal Microbalance (QCM) data

[5] Global Positioning System (GPS) dataset

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| GPS collection | 5,741 | fluorescence intensity data | 1 | academic test score data | 1 | MS [6]Word | 11,482 |
| text | 4,033 | MS Excel sheet | 1 | - | - | software | 8,176 |
| navigation primary | 3,691 | datasets containing results of materials testing and accompanying info. | 1 | - | - | PDF | 6,626 |
| protein sequence data | 2,803 | dataset | 4 | - | - | VND Excel | 3,495 |
| digital | 2,699 | - | 3,760 | - | - | tools | 2,428 |
| image | 1063 | - | | - | - | text plain | 686 |
| **totals** | **97,457** | **totals** | **7,520** | **totals** | **9,832** | **totals** | **230,721** |

[6] Microsoft

Table 16 and Table 17 display the top 10 repositories of published records in the STEM fields in the DCI. Data repositories have been studied because they function as the centralized infrastructure for research data, ensuring greater visibility for future reuse thanks to a readily available infrastructure for preservation of, access to, and reuse. Data repositories also play a crucial role in data sharing because papers for which micro data are available in a public repository received on average 9% more citations than those that did not make such data available (Piwowar & Vision, 2013). In order to examine closely how STEM repositories service data sharing, it is necessary to examine practices within research data repositories more generally; doing so will also reveal disciplinary differences regarding repositories.

Repositories – which can be housed within data centers or libraries - host and manage research data, playing a central role in data stewardship, accessibility, and persistence and facilitating conversion of metadata in to data. The findings indicate that data repositories were quite diverse across the scientific disciplines of the STEM fields. As can be seen, some digital repositories are much more widely used than others, depending on the disciplines that they serve. The findings further indicate that only a few data types comprised the published/shared research data in the DCI, once again except for astronomy/physics. The types of digital repositories in which data sharers can preserve their research data include commercial, institutional, governmental, and multidisciplinary repositories, and the websites of firms, journals and individuals. In general, sharers preferred to preserve their research data in third-party digital repositories rather than on the websites of journals.

As part of the analysis, typos or varying forms of terms for data types were merged (e.g., "EC-IRC Petten Institute for Energy and Transport", "EC-JRC Petten Institute for Energy and Transport", "EC JRC Petten Institute for Energy" and "EC-JRC Petten Institute for Energy" for

"EC JRC Petten IET[7]". Interestingly, a journal publisher's repository (i.e., that of the International Journal of Engineering and Science) has been used to preserve research data. Given that the data repositories of third parties or institutions are becoming widely used for data sharing, the repositories of journal publishers merit further study in this regard.

As just noted, and as can be seen in Table 16, institutions, associations, and governmental agencies are among the entities that maintain repositories. In astronomy/physics, the ten most used repositories were institutional. In the biological sciences, the most-used repositories were domain-specific or maintained by government agencies or associations. In the hard sciences, such as chemistry, institutional repositories (e.g., Cambridge Structural Database) were home to more than 50% of digital repository records; such repositories are widely observed as part of data-sharing infrastructure. In computing, sharers make use of company-specific repositories, such as Google's, as well as institution-specific repositories such as those maintained by the University of Washington and University of Athens. In the biological sciences, some generic repositories such as Dryad, are also observed, while others, such as Zenodo or Figshare, are not.

As Table 17 illustrates, both the earth sciences and mathematical sciences demonstrated unevenness in regard to data sharing, with PANGEA constituting 69.81% of data sharing in the former and CRAN 82.73% in the latter. Data citation therefore occurred at the repository -level rather than the data -level, as shown in Table 12. U.S. national repositories, such as those of NCDC and NOAA, were also used in the earth sciences.

Table 16 Repositories: Top 10 repositories for published records of the astronomy/physics, biological sciences, chemistry and computing fields in the DCI

[7] Institute for Energy and Transport

| **astronomy & physics** | | **biological sciences** | | **chemistry** | | **computing** | |
|---|---|---|---|---|---|---|---|
| **repository** | **total records** | **repository** | **total records** | **repository** | **total records** | **repository** | **total records** |
| EAWAG[8] | 7,509 | GEO[9] | 1,459,500 | CSD[10] | 490,251 | ISTI[11] | 1,122 |
| Keio[12] | 4,780 | UniProKB[13] | 535,810 | COD[14] | 329,875 | Gitter | 324 |
| UFZ[15] | 2,758 | Arrayexpress Archive | 417,716 | Pitt Quantum Repository | 106,059 | Univ. of Athens[16] | 295 |
| WSU[17] | 2,623 | ENA[18] | 225,122 | EMData Bank[19] | 5,394 | UW[20] | 264 |

[8] Swiss Federal Institute of Aquatic Science and Technology (EAWAG)

[9] Gene Expression Omnibus

[10] Cambridge Structural Database

[11] istituto di scienza e tecnologie dell informazione a faedo cnr

[12] Keio university

[13] Uniprot KnowledgeBase

[14] Crystallography Open Database

[15] Helmholtz Centre for Environmental Research Ufz Gmbh

[16] National Kapodistrian University of Athens

[17] Washington State University

[18] European Nucleotide Archive

[19] Electron Microscopy Data Bank

[20] University of Washington

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Univ. of Tokyo | 2,471 | YRC Public Image Repository[21] | 113,106 | GSDB[22] | 4,299 | Imperial College London | 237 |
| Nara Women s Univ. | 1,773 | wwPDB[23] | 106,038 | Chemotion | 4,047 | IBNL | 214 |
| RIKEN[24] | 1,719 | DEG-NCBI[25] | 106,025 | SDBS[26] | 3,720 | Univ. of Oxford | 191 |
| Univ. of Athens | 1,476 | Animal GTL Database | 82,640 | eCrystals[27] | 572 | Google | 185 |
| TUT[28] | 1,267 | PlantTFDB[29] | 65,536 | SBGrid Data Bank | 264 | CERN[30] | 174 |

[21] Yeast Resource Center Public Image Repository

[22] Bacterial Carbohydrate Structure Database

[23] Worldwide Protein Data Bank

[24]Kagaku Kenkyusho

[25] Database of Essential Genes - NCBI

[26] Spectral Database for Organic Compounds SDBS

[27] eCrystals – University of Southampton

[28] Toyohashi University of Technology

[29] Plant Transcription Factor Database

[30] European Organization for Nuclear Research

| | | | | | | | |
|---|--:|---|--:|---|--:|---|--:|
| Tohoku Univ. | 910 | Dryad | 65,329 | Tardis | 81 | Materials Virtual Lab | 156 |
| **totals** | **27,286** | **totals** | **3,176,822** | **totals** | **944,562** | **totals** | **3,162** |

Table 17 Repositories: Top 10 repositories for published records in the earth sciences, engineering, mathematical sciences and technology fields in the DCI

| **earth sciences** | | **engineering** | | **mathematical sciences** | | **technology** | |
|---|--:|---|--:|---|--:|---|--:|
| **repository** | **total records** | **repository** | **total records** | **repository** | **total records** | **repository** | **total records** |
| PANGAEA | 405,944 | EC IRC Petten IET[31] | 656 | CRAN[32] | 8,155 | Cell image library CCDB | 1,378 |
| SIOExplorer | 60,195 | MPI for Intelligent Systems[33] | 37 | Univ. of Florida Sparse Matrix | 1,640 | CDPH Merced District | 717 |

[31] European Commission - Institute for Energy and Transport

[32] Comprehensive R Archive Network

[33] Max Planck Institute for Intelligent Systems

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | Collection | | | |
| PISCO[34] | 29,230 | LEI[35] | 11 | J-Pal[36] | 52 | California Water Service | 653 |
| R2R[37] | 18,224 | NRG[38] | 11 | IGeS Database[39] | 35 | CDPH San Bernardino District | 647 |
| UC3 Merritt Repository[40] | 14,794 | NIMS[41] | 8 | 3TU Datacentrum | 17 | IJES[42] | 597 |

[34] Partnership for Interdisciplinary Studies of Coastal Oceans

[35] Lithuanian Energy Institute

[36] The Abdul Latif Jameel Poverty Action Lab

[37] Rolling Deck to Repository

[38] NRG Petten

[39] International Geoid Service Database

[40] University of California Curation Center (UC3) Merritt Repository

[41] National Research Institute for Metals

[42] International Journal of Engineering and Science

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| IEDA[43] | 14,351 | OECD[44] | 4 | - | - | CDPH Sonoma District | 568 |
| NODC[45] | 9,155 | - | - | - | - | CDPH Tehachapi District | 420 |
| NOAA Paleoclimatology[46] | 9,109 | - | - | - | - | San Bernardino County | 406 |
| UNAVCO[47] | 3,992 | - | - | - | - | Fresno County | 396 |
| NOAA | 2,974 | - | - | - | - | CDPH Klamath District | 375 |
| **totals** | **101,829** | **totals** | **727** | **totals** | **9,899** | **totals** | **6,157** |

Table 18 and Table 19 display the top 10 most shared/published data authors for each of the STEM disciplines in the DCI. In engineering, two data authors contributed each 33.01% and

[43] Interdisciplinary Earth Data Science - Marine Geoscience Data System
[44] Organization for Economic Cooperation and Development OECD
[45] US National Oceanographic Data Center
[46] National Oceanic and Atmospheric Administration (NOAA) Paleoclimatology
[47] UNAVCO Geodesy Data Archive

31.47%, respectively of the data publications. In other disciplines, such as the earth sciences, mathematical sciences and technology, by contrast, no author contributed more than 5% of total data publications.

In any care, a relatively small number of data authors tend to contribute a large proportion of data publications. High rates of data sharing among just a few authors in a discipline (e.g., three authors being responsible for more than 30% of sharing within a single discipline) tend not to be observed in standard bibliographic publications, such as in the journal articles, books, or conference proceedings. Researchers may be pioneers in this regard. The relatively high rates of anonymous sharing in astronomy (5.66%) and the biological sciences (3.18%) could complicate the rewarding of formal scholarly credit for data sharers.

Table 18 Data authors: The top 10 most highly shared/published data authors of the astronomy/physics, biological sciences, chemistry and computing fields in the DCI

| **astronomy/physics** | | **biological sciences** | | **chemistry** | | **computing** | |
|---|---|---|---|---|---|---|---|
| **data authors** | **total records** | **data authors** | **total records** | **data authors** | **total records** | **data authors** | **total records** |
| Schymanski E | 7,509 | anonymous | 116,017 | Anonymous | 5,602 | Howison James | 1,713 |
| Singer H | 7,509 | Shah P | 72,099 | Hursthouse Michael B | 2,909 | Crowston Kevin | 1,712 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Stravs M | 7,509 | Sherlock G | 72,098 | Fun Hoong-Kun | 2,819 | Squire Megan | 1,711 |
| Horai H | 4,780 | Binkley J | 72,094 | Ng Seik Weng | 2,585 | Assante Massimiliano | 368 |
| Kakazu Y | 4,780 | Binkley G | 72,093 | Rheingold Arnold L | 2,138 | Perciante Costantino | 249 |
| anonymous | 3,915 | Inglis Do | 72,093 | Zhang Yong | 2,114 | anonymous | 244 |
| Berger S | 3,171 | Miyasato Sr | 72,093 | Ng SW | 2,004 | Panichi Giancarlo | 224 |
| Braun S | 3,167 | Simison M | 72,093 | Jones PG | 1,989 | Sinibaldi Fabio | 210 |
| Kalinowski H-O | 3167 | Skrzypek MS | 72,093 | Skelton BW | 1,974 | Ong Shyue Ping | 209 |
| Schulze T | 2758 | Wymore F | 72,093 | Ma Jian-Fang | 1,925 | Perez Jose | 161 |
| **totals** | **45,098** | **totals** | **746,866** | **totals** | **26,059** | **totals** | **6,801** |

Table 19 Data authors: The top 10 most highly shared/published data authors for earth sciences, engineering, mathematical sciences and technology in the DCI

| **earth sciences** | | **engineering** | | **mathematical sciences** | | **technology** | |
|---|---|---|---|---|---|---|---|
| **data authors** | **total records** | **data authors** | **total records** | **data authors** | **total records** | **data authors** | **total records** |
| Hofmann Jutta | 19,569 | Ennis PJ | 2,355 | Hoekstra R | 192 | Rzepa Henry S | 2,046 |
| Konig-Langlo Gert | 1,4785 | Offerman M | 2,245 | Welker V | 128 | Wang Wei | 1,586 |
| Schweingruber Fritz Hans | 1,3003 | Basan Robert | 1,035 (14.51%) | Rao A | 91 (0.92%) | Zhang Wei | 1,370 |
| Bleyer Hans-Jurgen | 9,131 | Marohnic Tea | 1,035 | Senses B | 91 | Li Yan | 1,070 |
| Washburn Libe | 8,774 | Mccolvin GM | 1,007 | Banerjee Abhijit | 52 | Wang Jun | 1,032 |
| Menge Bruce | 8,483 | Rubesa Domagoj | 833 | Cole Shawn | 52 | Wang Ying | 1007 |

| | | | | | | | |
|---|---:|---|---:|---|---:|---|---:|
| Chan Francis | 8,298 | Mccarthy S | 583 | Duflo Esther | 52 | Baldock Richard | 956 |
| Mcmanus Margaret | 7,218 | Schneider K | 244 | Linden Leigh | 52 | Wang Jing | 952 |
| Friedrich Michael | 6428 | de Haan F | 227 | Zhao W | 50 | Richardson Lorna | 947 |
| Zuyev Aleksey N | 5549 | Papuga Jan | 202 | Chamberlain Scott | 44 | Zhang Yan | 945 |
| **totals** | **101,238** | **totals** | **9,766** | **totals** | **752** | **totals** | **11,911** |

Table 20 and Table 21 display the top 10 most highly shared/published author groups in STEM fields in the DCI. I included author groups in this study because modern research frequently involves collaboration among multiple labs, departments and institutions whether within a region or internationally. The typos “JMC OICR” and “JMG-OICR” as were engineering typos or different uses of terms for data types.

In the case of four disciplines, no one group of authors was dominant among the rest, notable contributors were observed in the biological sciences (UniProt, 14.49%), chemistry (PITT Quantum Repository, 11.23%), the earth sciences (MEDAR Group, 16.84%), and engineering (EC JRC Petten IET, 8.56%).

Table 20 Group authors: The top 10 most highly shared/published group authors for astronomy/physics, biological sciences, chemistry and computing in the DCI

| **astronomy/physics** | | **biological sciences** | | **chemistry** | | **computing** | |
|---|---|---|---|---|---|---|---|
| **group authors** | **total records** | **group authors** | **total records** | **group authors** | **total records** | **group authors** | **total records** |
| MSSJ[48] | 1,959 | UniProt | 528,670 | PITT Quantum Repository | 106,058 | The Gitter Badger | 270 |
| Soda Aromatic Co Ltd | 1,039 | PISCO | 32,483 | DLUT[49] | 6 | Mosquito Alert | 67 |
| Complete Team | 572 | EcoTrends[50] | 30,806 | State Key Laboratory of Supramolecular Structure and Materials | 6 | Making GitHub Delicious | 60 |

[48] Mass Spectroscopy Society of Japan

[49] Dalian University of Technology

[50] Ecotrends Project

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| Lambda NASA GSFC | 528 | miRBase | 26,717 | CT | 2 | CedergroupClusters | 56 |
| Kuraray Co Ltd | 421 | JGI[51] | 20,911 | - | - | Cmsbuild | 48 |
| CMS Collaboration | 408 | USGS[52] | 18,690 | - | - | Yanikou19 | 39 |
| Nara Women's Univ. | 223 | Broad Institute | 17,137 | - | - | Flxb | 35 |
| Ube Scientific Analysis Laboratory | 137 | encode dcc | 11,370 | - | - | JMG OICR | 58 |
| School of Medicine Hamamatsu Univ. | 104 | - | - | - | - | rgmumufeng | 27 |

[51] DOE Joint Genome Institute

[52] United States Geological Survey

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ISIR, Osaka Univ.[53] | 100 | - | - | - | - | - | - |
| **totals** | **5,491** | **totals** | **686,784** | **totals** | **106,072** | **totals** | **660** |

Table 21 Group authors: The top 10 most highly shared/published group authors for earth sciences, engineering, mathematical sciences and technology in the DCI

| **earth sciences** | | **engineering** | | **mathematical sciences** | | **technology** | |
|---|---|---|---|---|---|---|---|
| **group authors** | **total records** | **group authors** | **total records** | **group authors** | **total records** | **group authors** | **total records** |
| MEDAR group | 97,899 | EC JRC Petten IET | 636 | IBM | 69 | CellimageliBrary CCDB | 1,378 |
| PISCO | 27,703 | Max Planck Institute for Metallforschung | 37 | RAJAT | 31 | CDPH Merced District | 717 |

[53] Institute of Scientific and Industry Research, Osaka University

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| GDC[54] | 26,213 | LEI55 | 11 | NASA[56] | 22 | California Water Service | 653 |
| WOCE Sea Level WSL | 19,649 | NRG PETTEN | 11 | Integrated Sys Eng | 18 | CDPH San Bernardino District | 647 |
| R2R[57] Program | 18,212 | National Research Institute for Metals | 8 | Autoform Eng | 16 | The IJES | 597 |
| Shipboard Scientific Party | 17,530 | OECD[58] | 4 | COMSOL[59] | 9 | CDPH Sonoma District | 568 |

[54] Geological Data Center
[55] Lithuanian Energy Institute
[56] National Aeronautics and Space Administration
[57] Rolling Deck to Repository
[58] Organization for Economic Co-operation and Development
[59] COMSOL Multiphysics

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| WOCE WHP[60] | 16,946 | - | - | Geofarik GmbH | 8 | CDPH Tehachapi District | 420 |
| WOCE UOT[61] | 12,894 | - | - | Francesca Petralia Developer | 2 | San Bernardino County | 406 |
| OMEX Project Members | 6,006 | - | - | Graph Drawing Contest | 2 | Fresno County | 396 |
| WOCE SVP[62] | 5,921 | - | - | KDD cup 2003 | 2 | CDPH Klamath District | 375 |
| **totals** | **248,973** | **totals** | **727** | **totals** | **179** | **totals** | **6,157** |

To summarize, RQ1 addressed the diversity of data sharing practices as measured by data citation across STEM variables. In addition to this question, the following trends were observed.

[60] WOCE Hydrographic Program
[61] WOCE Upper Ocean Thermal UOT
[62] WOCE Surface Velocity Program SVP

- Each STEM field had distinct data sharing practices (i.e., these practices are characterized by unevenness), and the distribution of data sharing was very skewed. Thus, for instance, data sharing was most prevalent in the biological sciences.
- Funding agencies' data sharing requirements served as major triggers for data sharing in all of the STEM fields. Thus, for example, the total instances of data sharing rose dramatically after 2003, when the first data sharing mandate was implemented.
- Diverse types of digital repositories were observed across STEM disciplines, but one type usually dominated in a given discipline. Thus, for instance, governmental agencies maintained the repositories most used in the biological sciences, while in the earth sciences, the discipline-specific repository PANGEA was the locus of 69.81% of data sharing, and in the mathematical sciences the discipline-independent repository CRAN was the locus of 82.73% of data sharing.

### 4.2.2. RQ2: What types of STEM research data are formally cited most often?

RQ2 was answered using descriptive analysis and the results reported based on the DCI records at the data level.

- Objective of RQ2: To examine the types of research data most often cited formally.

Table 22 lists the data types that are most often cited in STEM fields in the DCI. Topping the list in terms of frequency is "data file," which was the form of 89,934 (1.76%) of citations. Next, in order, were sequence data (74,642 total times cited, 1.46%), crystallographic data (52,078 total times cited, 1.02%), blank meaning no data type w displayed (43,375 total times cited, 0.85%), software code (41,634 total times cited, 0.81%), mass spectral data (31,072 total times cited, 0.61%), crystal structure (29,209 total times cited, 0.57%), molecular structure (11,144 total times cited, 0.22%), sequence read archive (8,979 total times cited, 0.18%), fileset (8,229 total times cited, 0.16%) and nuclear magnetic resonance (6,157 total times cited, 0.12%).

Quantitative data were more often cited and shared than qualitative data because the STEM fields emphasize this kind of data. The sole example qualitative research data observed was in technology in 2017. It took the form of a transcript of an interview, and thus represented the interviewee's own words (see discussion below).

Table 22 displays the top 10 types of data that received the most citations in STEM fields in the DCI, none of which was dominant secondary data types, such as MS Excel sheets, MS Word, MS PowerPoint, or digital video files, were not observed, and in many cases the data type was not recorded (i.e., fields were left blank). Improvements in data curation could thus include (1) modification of data structures in the DCI and (2) optional fields for data types in records maintained by data-sharing repositories. Given the scope and limitations of the datasets used in the present study, an examination of the reasons is left for future research.

Table 22 Top 10 data types that received the most data citation in STEM fields in the DCI

| **data type** | **total times cited** | **percentage** |
|---|---|---|
| data file | 89,934 | 1.76% |
| protein sequence data | 74,642 | 1.46% |
| crystallographic data | 52,078 | 1.02% |
| (blank) | 43,375 | 0.85% |
| software code | 41,634 | 0.81% |
| mass spectral data | 31,072 | 0.61% |

| | | |
|---|---|---|
| crystal structure | 29,209 | 0.57% |
| molecular structure | 11,144 | 0.22% |
| sequence read archive | 8,979 | 0.18% |
| fileset | 8,229 | 0.16% |
| nuclear magnetic resonance results | 6,157 | 0.12% |
| **totals** | **475,803** | **9.31%** |

Table 23 to Table 30 display detailed examinations of disciplinary differences among the top 10 data types that are most cited data types for each discipline. These differences also merit further study, as many aspects of data sharing are discipline-specific. The total numbers of data types varied across disciplines owing to the diversity of data sharing practice, as seen in relation to RQ1.

Table 23 displays the top 10 most highly cited data types in the DCI in astronomy/physics. The distribution of data type in astronomy/physics was quite skewed in certain data types. These top 10 data types accounted for 98.62% of citations, and the top three (mass spectral data, NMR results, and spectral data) for over 90% in this discipline

Table 23 Astronomy/physics: Top 10 most highly cited data types

| **data type** | **total times cited** | **percentage** |
|---|---|---|
| mass spectral data | 31,072 | 70.80% |
| NMR results | 6,157 | 14.03% |
| spectral data | 3,723 | 8.48% |

| | | |
|---|---|---|
| software | 1,396 | 3.18% |
| image file | 234 | 0.53% |
| FITS file | 192 | 0.44% |
| data/dataset | 170 | 0.39% |
| final output picture | 163 | 0.37% |
| dataset | 63 | 0.14% |
| HRCROP | 60 | 0.14% |
| TEX APPB | 50 | 0.11% |
| **totals** | **43,280** | **98.62%** |

Table 24 displays the top 10 most highly cited data types in the DCI in the biological sciences. The top three (RNA, protein sequence data, and SRA) accounted for more than half of the citations.

Table 24 Biological sciences: Top 10 most highly cited data types

| data type | total times cited | percentage |
|---|---|---|
| RNA[63] | 931,673 | 29.67% |
| protein sequence data | 525,973 | 16.75% |
| SRA[64] | 277,920 | 8.85% |
| Genomic | 163,349 | 5.20% |
| images | 113,107 | 3.60% |
| nucleotide sequencing information | 109,135 | 3.48% |

[63] Ribonucleic acid
[64] Sequence Read Archive

| | | |
|---|---:|---:|
| molecular structure | 75,899 | 2.42% |
| FGEM | 72,717 | 2.32% |
| Processed | 72,717 | 2.32% |
| plant transcription factors and their annotation | 65,536 | 2.09% |
| **totals** | **2,408,026** | **76.69%** |

Table 25 displays the top 10 most highly cited data types in the DCI in chemistry. The top 10 data types accounted for 99.95% which indicates the existence of major data types to be cited in chemistry. Most notably, the top two data types (crystal structure, crystallographic data) account for 86.61% of the citations.

Table 25 Chemistry: Top 10 most highly cited data types

| **data type** | **total times cited** | **percentage** |
|---|---:|---:|
| crystal structure | 754,913 | 52.51% |
| crystallographic data | 490,252 | 34.10% |
| molecular structure | 91,870 | 6.39% |
| crystallographic information | 84,687 | 5.89% |
| bacterial carbohydrate structure | 4,298 | 0.30% |
| spectral data | 3,720 | 0.26% |
| crystallographic structure | 3,008 | 0.21% |
| dataset | 2,410 | 0.17% |
| molecular data | 954 | 0.07% |

| | | |
|---|---:|---:|
| molecule | 647 | 0.05% |
| **totals** | **1,436759** | **99.95%** |

Table 26 displays the top 10 most highly cited data types in the DCI in computing. Not surprisingly, software comprised over 91.41% (18,246 total times cited) of citations.

Intuitively, the low percentage of total times cited counts in computing is remarkable given the prevalence of software code in the discipline compared with others. This finding is explicable in terms of the usage of proprietary software in computing.

Table 26 Computing: Top 10 most highly cited data types

| **data type** | **total times cited** | **percentage** |
|---|---:|---:|
| software | 18,246 | 91.41% |
| code | 1,278 | 6.40% |
| model | 416 | 2.08% |
| dataset | 3 | 0.02% |
| database | 2 | 0.01% |
| other | 2 | 0.01% |
| raw experimental data | 2 | 0.01% |
| chemistry data | 1 | 0% |
| dataset used in the paper | 1 | 0% |
| diagrams | 1 | 0% |
| **totals** | **19,952** | **99.95%** |

Table 27 displays the top 10 most highly cited data types in the DCI in the earth sciences. In this discipline, geospatial datasets such as GPS datasets, which can be reused to visualize spatiotemporal analyses based on the computational use of source code, are identified as highly cited datasets, accounting for 30.64% of citations.

Table 27 Earth sciences: Top 10 most highly cited data types

| **data type** | **total times cited** | **percentage** |
|---|---:|---:|
| dataset | 32,975 | 30.64% |
| interactive resource | 22,264 | 20.69% |
| GPS dataset | 13,080 | 12.15% |
| geoscientific information | 9,108 | 8.46% |
| GPS collection | 5,741 | 5.33% |
| text | 4,033 | 3.75% |
| navigation primary | 3,691 | 3.43% |
| protein sequence data | 2,803 | 2.60% |
| digital | 2,699 | 2.51% |
| **totals** | **96,394** | **89.56%** |

Table 28 displays the top 10 most highly cited data types in the DCI in engineering. In this discipline, seven data types accounted for 100% of the records, though test data alone accounted for 99.71% of the citations, followed distinctly by datasets (0.13%).

Table 28 Engineering: Top 10 most highly cited data types

| data type | total times cited | percentage |
|---|---:|---:|
| test data | 3,749 | 99.71% |
| dataset | 5 | 0.13% |
| GIS vector data | 2 | 0.05% |
| QCM [65]data | 1 | 0.03% |
| microscopy images | 1 | 0.03% |
| fluorescence intensity data | 1 | 0.03% |
| MS Excel spreadsheet | 1 | 0.03% |
| **totals** | **3,760** | **100%** |

Table 29 displays the top 10 most highly cited data types in the mathematical sciences. 'GEOID ondulation given on a grid' was combined with 'GEOID undulation given on a grid' because ondulation is a typo. In this discipline, five data types accounted for 100% of the data types cited in the DCI the top two, software (82.94%) and matrix (16.69%), accounting for 99.63% of citations.

Table 29 Mathematical sciences: Top 10 most highly cited data types

| data type | total times cited | percentage |
|---|---:|---:|
| software | 8,155 | 82.94% |
| Matrix | 1,640 | 16.69% |
| GEOID undulation given on a grid | 35 | 0.35% |

[65] Quartz Crystal Microbalance

| | | |
|---|---:|---:|
| dataset | 1 | 0.01% |
| academic test score data | 1 | 0.01% |
| **total** | **9,832** | **100%** |

Table 30 displays the top 10 most highly cited data types in technology (N=235,315). In technology, the top 10 most highly cited data types accounted for 98.05% of the all data types. The most highly cited type were datasets (58.38%), followed by filesets (14.15%).

In light of these observations, consideration needs to be given to standards for text files. Thus, for instance, if a data file is a PDF, MS Excel spreadsheet, or a graph or table, searchability for reuse is a concern.

Table 30 Technology: Top 10 most highly cited data types

| **data type** | **total times cited** | **percentage** |
|---|---:|---:|
| dataset | 137,375 | 58.38% |
| fileset | 33,304 | 14.15% |
| image TIFF [66] | 14,558 | 6.19% |
| image | 12,591 | 5.36% |
| application MS Word | 11,482 | 4.88% |
| software | 8,176 | 3.47% |
| application PDF [67] | 6,626 | 2.82% |
| application VND MS Excel | 3,495 | 1.49% |

[66] Tagged Image File Format

[67] Portable Data Format

| | | |
|---|---:|---:|
| tools | 2,428 | 1.03% |
| text plain | 686 | 0.29% |
| **totals** | **230,721** | **98.05%** |

To summarize the results for RQ2:

- The data types cited were very diverse in STEM disciplines.
- Nearly all of the data cited in STEM disciplines were quantitative in nature; only a single example of qualitative data, an interview transcript, was observed.
- The top 10 most highly cited data types in STEM disciplines were, in descending order, data files, protein sequence data, crystallographic data, n/a (i.e., no data type was specified by data sharers), software code, mass spectral data, crystal structure, molecular structure, sequence read archive, filesets, nuclear magnetic resonance and nuclear magnetic resonance results.
- The data types that most often cited varied across STEM disciplines.

### 4.2.3. RQ3: How do author self-citation/recitation practices differ across STEM disciplines?

- Objective 1: To identify factors associated with author self-citation and recitation.
- Objective 2: To examine these factors (across and within discipline).

Table 31 displays the comparative analysis of self-citation at the data-level in the DCI using citer-based analysis. WoS databases were used for all of the article-level data except for the DCI. Low levels of author self-citation and recitation were observed, on average (3.91%), and slight differences between the data and article-level outcomes; thus, the average author self-citation rate was 3.94% at the data-level and 3.88% at the article-level. The greatest difference between data and article-level outcomes was in computing (1.68% difference) and the least difference was with mathematical sciences (0.03% difference). At the article-level, a relatively high author self-citation rate was observed in computing, the earth sciences, and technology.

Table 31 Comparisons of self-citation between data-level and article-level by citer-based analysis

| **discipline** | **no. of authors** | **data-level** | | | **article-level** | | |
|---|---|---|---|---|---|---|---|
| | | **sum of times cited** | **without self-citations** | **self-citations** | **sum of times cited** | **without self-citations** | **self-citations** |
| **astronomy & physics** | 23 | 37654 | 35115 | 2539 (6.74%) | 37239 | 34822 | 2417 (6.49%) |
| **biological sciences** | 29 | 302522 | 292701 | 9821 (3.25%) | 299259 | 289733 | 9526 (3.18%) |
| **chemistry** | 25 | 65111 | 60366 | 4745 (7.29%) | 63918 | 59493 | 4425 (6.92%) |
| **computing** | 28 | 7305 | 7274 | 31 (0.42%) | 7063 | 6915 | 148 (2.1%) |
| **earth sciences** | 24 | 20695 | 19911 | 784 (3.79%) | 15498 | 14811 | 687 (4.43%) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **engineering** | 22 | 5282 | 5040 | 242 (4.58%) | 5279 | 5039 | 240 (4.55%) |
| **mathematical sciences** | 29 | 17693 | 17356 | 337 (1.9%) | 17549 | 17224 | 325 (1.85%) |
| **technology** | 25 | 19269 | 19014 | 255 (1.32%) | 19254 | 18989 | 265 (1.38%) |
| **average** | **205** | **475531** | **456777** | **18754 (3.94%)** | **465059** | **447026** | **18033 (3.88%)** |

Table 32 to 37 display outcomes of the Kruskal-Wallis tests. Table 32 to 34 display at the data-level. Table 35 to 37 display at the article-level. The groups (i.e., from group 1 to group 8) differed in terms of numbers of members (i.e., authors). It was, as mentioned above, owing to this violation of the one-way ANOVA assumption (i.e., same group numbers) that the Kruskal-Wallis test was conducted as a means to examine the associations within and across shared research data and the instances of author self-citation or recitation in the various STEM fields in greater detail. The numbers assigned to each group were as follows: group 1 for astronomy/physics, group 2 for the biological sciences, group 3 for chemistry, group 4 for computing, group 5 for the earth sciences, group 6 for engineering, group 7 for the mathematical sciences and group 8 for technology.

Table 32 Data-level: Self-citation rate descriptive statistics for each discipline

**Descriptives**

DataSelfCiteRate

| | | | | | 95% Confidence Interval for Mean | | | |
|---|---|---|---|---|---|---|---|---|
| | N | Mean | Std. Deviation | Std. Error | Lower Bound | Upper Bound | Minimum | Maximum |
| 1 | 23 | .0478 | .04123 | .00860 | .0300 | .0657 | .00 | .14 |
| 2 | 29 | .0362 | .02611 | .00485 | .0263 | .0461 | .00 | .10 |
| 3 | 25 | .0836 | .06915 | .01383 | .0551 | .1121 | .00 | .28 |
| 4 | 28 | .0032 | .01249 | .00236 | -.0016 | .0081 | .00 | .06 |
| 5 | 24 | .0421 | .04501 | .00919 | .0231 | .0611 | .00 | .18 |
| 6 | 22 | .0736 | .08533 | .01819 | .0358 | .1115 | .00 | .32 |
| 7 | 29 | .0197 | .02612 | .00485 | .0097 | .0296 | .00 | .09 |
| 8 | 25 | .0120 | .01732 | .00346 | .0049 | .0191 | .00 | .06 |
| Total | 205 | .0382 | .05166 | .00361 | .0311 | .0453 | .00 | .32 |

Table 33 displays the independent-samples Kruskal-Wallis test to examine the author self-citation rate across disciplines. In light of these results, one author in the discipline of chemistry was removed as an outlier, because, as noted above, his last name, Lee is very common in East Asian countries

Table 33 Data-level: Independence-samples Kruskal-Wallis test. (Self-Citation-Rate across discipline)

**Independent-Samples Kruskal-Wallis Test Summary**

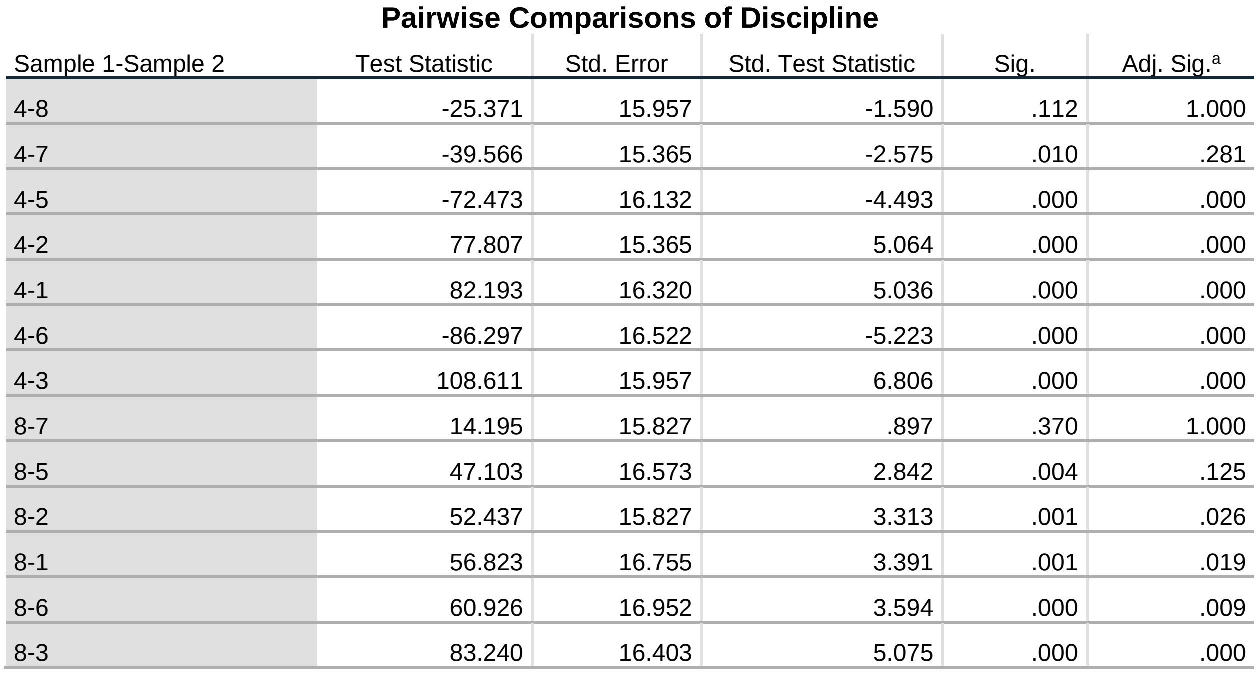
| | |
|---|---|
| Total N | 205 |
| Test Statistic | 71.790[a] |
| Degree Of Freedom | 7 |
| Asymptotic Sig.(2-sided test) | .000 |

a. The test statistic is adjusted for ties.

Table 34 displays the pairwise comparisons of disciplines using the independent-samples Kruskal-Wallis test at the data-level. The distribution of DataLevelSelfCiteRate (i.e., author self-citation) at the data-level differed across disciplines. The null hypothesis ($H_0$ : the distribution of DataLevelSelfCiteRate is the same across categories of Discipline) was rejected because the p value of .000 was less than 0. 01 (p <0. 01).

The p-values for the Kruskal-Wallis test comparing the author self-citation rate and disciplines showed significant differences ( $\rho$ < 0.01) for several groups from the multiple pairwise comparisons by discipline regarding DataLevelSelfCitationRate at the data-level. The pairs of groups were 4-5 (computing – earth sciences), 4-1 (computing – astronomy/physics), 4-2 (computing – biological sciences), 4-6 (computing – engineering),4-3 (computing – chemistry), 8-3 (technology – chemistry), and 7-3 (mathematical sciences-chemistry). There was no evidence of such differences between the other groups.

Table 34 Data-level: Pairwise comparisons of discipline (Kruskal-Wallis test)



**Pairwise Comparisons of Discipline**

| Sample 1-Sample 2 | Test Statistic | Std. Error | Std. Test Statistic | Sig. | Adj. Sig.[a] |
|---|---|---|---|---|---|
| 4-8 | -25.371 | 15.957 | -1.590 | .112 | 1.000 |
| 4-7 | -39.566 | 15.365 | -2.575 | .010 | .281 |
| 4-5 | -72.473 | 16.132 | -4.493 | .000 | .000 |
| 4-2 | 77.807 | 15.365 | 5.064 | .000 | .000 |
| 4-1 | 82.193 | 16.320 | 5.036 | .000 | .000 |
| 4-6 | -86.297 | 16.522 | -5.223 | .000 | .000 |
| 4-3 | 108.611 | 15.957 | 6.806 | .000 | .000 |
| 8-7 | 14.195 | 15.827 | .897 | .370 | 1.000 |
| 8-5 | 47.103 | 16.573 | 2.842 | .004 | .125 |
| 8-2 | 52.437 | 15.827 | 3.313 | .001 | .026 |
| 8-1 | 56.823 | 16.755 | 3.391 | .001 | .019 |
| 8-6 | 60.926 | 16.952 | 3.594 | .000 | .009 |
| 8-3 | 83.240 | 16.403 | 5.075 | .000 | .000 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 7-5 | 32.907 | 16.003 | 2.056 | .040 | 1.000 |
| 7-2 | 38.241 | 15.229 | 2.511 | .012 | .337 |
| 7-1 | 42.627 | 16.192 | 2.633 | .008 | .237 |
| 7-6 | 46.731 | 16.396 | 2.850 | .004 | .122 |
| 7-3 | 69.045 | 15.827 | 4.363 | .000 | .000 |
| 5-2 | 5.334 | 16.003 | .333 | .739 | 1.000 |
| 5-1 | 9.720 | 16.922 | .574 | .566 | 1.000 |
| 5-6 | -13.824 | 17.117 | -.808 | .419 | 1.000 |
| 5-3 | 36.138 | 16.573 | 2.181 | .029 | .818 |
| 2-1 | 4.386 | 16.192 | .271 | .786 | 1.000 |
| 2-6 | -8.490 | 16.396 | -.518 | .605 | 1.000 |
| 2-3 | -30.803 | 15.827 | -1.946 | .052 | 1.000 |
| 1-6 | -4.104 | 17.294 | -.237 | .812 | 1.000 |
| 1-3 | -26.417 | 16.755 | -1.577 | .115 | 1.000 |
| 6-3 | 22.314 | 16.952 | 1.316 | .188 | 1.000 |

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.

Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Table 35 to 37 display the article-level author self-citation outcomes. The dependent variable was the article-level author self-citation rate. Group 6 (engineering) had the highest mean values and group 7 (mathematical sciences) the lowest.

Table 35 Article-level: Self-citation rate descriptive statistics for each discipline

**Descriptives**

ArticleSelfCiteRate

|  |  |  |  |  | 95% Confidence Interval for Mean |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | N | Mean | Std. Deviation | Std. Error | Lower Bound | Upper Bound | Minimum | Maximum |
| 1 | 23 | .0474 | .04002 | .00835 | .0301 | .0647 | .00 | .14 |
| 2 | 29 | .0352 | .02444 | .00454 | .0259 | .0445 | .00 | .10 |
| 3 | 25 | .0808 | .06608 | .01322 | .0535 | .1081 | .00 | .26 |
| 4 | 28 | .0032 | .01249 | .00236 | -.0016 | .0081 | .00 | .06 |
| 5 | 24 | .0396 | .04298 | .00877 | .0214 | .0577 | .00 | .17 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 6 | 22 | .0732 | .08566 | .01826 | .0352 | .1112 | .00 | .32 |
| 7 | 29 | .0197 | .02612 | .00485 | .0097 | .0296 | .00 | .09 |
| 8 | 25 | .0200 | .04509 | .00902 | .0014 | .0386 | .00 | .22 |
| Total | 205 | .0383 | .05204 | .00363 | .0311 | .0455 | .00 | .32 |

Table 36 displays the results of the independent-samples Kruskal-Wallis test used to examine the author self-citation rate across disciplines. Based on the results, thirty-five authors were removed, leaving 205 to be analyzed.

Table 36 Article-level: Independence-sample Kruskal-Wallis test (Self-Citation-Rate across discipline)

**Independent-Samples Kruskal-Wallis Test Summary**

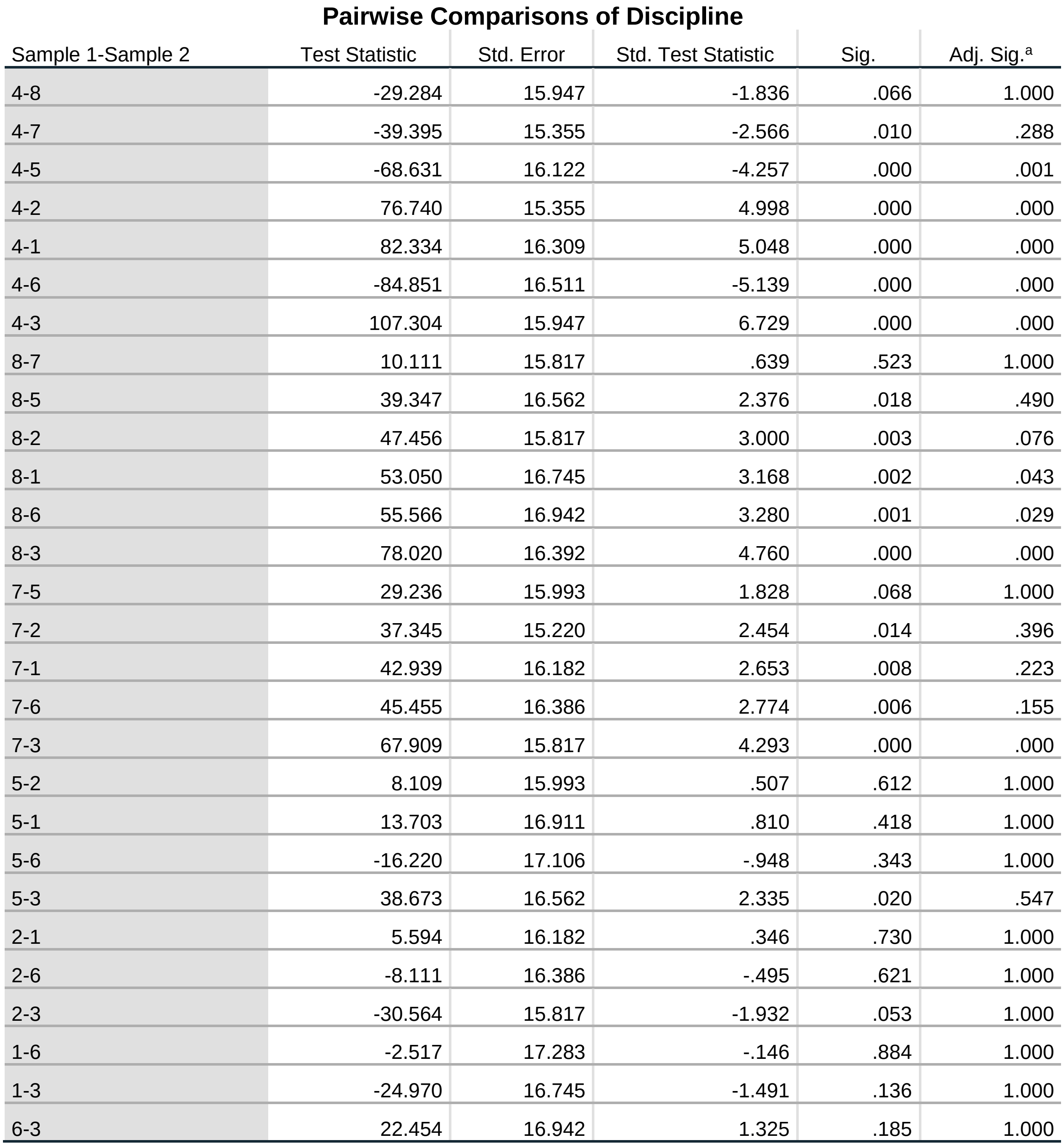
| | |
|---|---|
| Total N | 205 |
| Test Statistic | 67.749[a] |
| Degree Of Freedom | 7 |
| Asymptotic Sig.(2-sided test) | .000 |

a. The test statistic is adjusted for ties.

Table 37 displays the pairwise comparisons of disciplines using the independent-samples Kruskal-Wallis at the article-level. The distribution of ArticleLevelSelfCiteRate (i.e., author self-citations) at this level differed across disciplines. Based on these results, the null hypothesis ($H_0$ : the distribution of ArticleLevelSelfCiteRate is the same across disciplinary categories) as rejected because, again, the p value, 000, was less than 0.01 (p <0.01).

Several groups from the multiple pairwise comparison tests by discipline showed significant differences (p < 0.01) for ArticleLevelSelfCitationRate at the article-level. Thus, the difference between data and article-level was 4-6 (computing – engineering), which was added at the article-level. Groups showing significant differences were 4-5 (computing – earth sciences), 4-1 (computing – astronomy/physics), 4-2 (computing – biological sciences), 4-6 (computing – engineering),4-3 (computing – chemistry), 4-6 (computing – engineering), 8-3 (technology – chemistry), and 7-3 (mathematical sciences – chemistry).

Table 37 Article-level: Pairwise comparisons of discipline (Kruskal-Wallis test)



**Pairwise Comparisons of Discipline**

| Sample 1-Sample 2 | Test Statistic | Std. Error | Std. Test Statistic | Sig. | Adj. Sig.[a] |
|---|---|---|---|---|---|
| 4-8 | -29.284 | 15.947 | -1.836 | .066 | 1.000 |
| 4-7 | -39.395 | 15.355 | -2.566 | .010 | .288 |
| 4-5 | -68.631 | 16.122 | -4.257 | .000 | .001 |
| 4-2 | 76.740 | 15.355 | 4.998 | .000 | .000 |
| 4-1 | 82.334 | 16.309 | 5.048 | .000 | .000 |
| 4-6 | -84.851 | 16.511 | -5.139 | .000 | .000 |
| 4-3 | 107.304 | 15.947 | 6.729 | .000 | .000 |
| 8-7 | 10.111 | 15.817 | .639 | .523 | 1.000 |
| 8-5 | 39.347 | 16.562 | 2.376 | .018 | .490 |
| 8-2 | 47.456 | 15.817 | 3.000 | .003 | .076 |
| 8-1 | 53.050 | 16.745 | 3.168 | .002 | .043 |
| 8-6 | 55.566 | 16.942 | 3.280 | .001 | .029 |
| 8-3 | 78.020 | 16.392 | 4.760 | .000 | .000 |
| 7-5 | 29.236 | 15.993 | 1.828 | .068 | 1.000 |
| 7-2 | 37.345 | 15.220 | 2.454 | .014 | .396 |
| 7-1 | 42.939 | 16.182 | 2.653 | .008 | .223 |
| 7-6 | 45.455 | 16.386 | 2.774 | .006 | .155 |
| 7-3 | 67.909 | 15.817 | 4.293 | .000 | .000 |
| 5-2 | 8.109 | 15.993 | .507 | .612 | 1.000 |
| 5-1 | 13.703 | 16.911 | .810 | .418 | 1.000 |
| 5-6 | -16.220 | 17.106 | -.948 | .343 | 1.000 |
| 5-3 | 38.673 | 16.562 | 2.335 | .020 | .547 |
| 2-1 | 5.594 | 16.182 | .346 | .730 | 1.000 |
| 2-6 | -8.111 | 16.386 | -.495 | .621 | 1.000 |
| 2-3 | -30.564 | 15.817 | -1.932 | .053 | 1.000 |
| 1-6 | -2.517 | 17.283 | -.146 | .884 | 1.000 |
| 1-3 | -24.970 | 16.745 | -1.491 | .136 | 1.000 |
| 6-3 | 22.454 | 16.942 | 1.325 | .185 | 1.000 |

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.
a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

To summarize the results relating to RQ 3:

- The author self-citation and recitation rates were very low (i.e., 4% average) in STEM fields.
- The author self-citation, including recitation, rates differed slightly between the data-level (i.e., research data) and bibliographic-level (i.e., articles).
- Associations between and across shared research data and author self-citation and recitation were observed at the article level. Certain disciplines showed significant differences regarding author self-citation rates. The same groups with significant differences at both the data and article-level were (1) computing – earth sciences, (2) computing – astronomy/physics, (3) computing – engineering and (4) computing – chemistry. A difference between data-level and article-level associations was observed for computing – engineering. Two groups displaying significant differences at the article-level and not included at the data-level were computing – engineering.

### 4.2.4. RQ4: How do data reuse practices differ across STEM disciplines?

RQ4 was formulated to address the actual data reuse practices in data citation at the article- and discipline-levels. I combined automatic text-searching techniques with manual content analysis that involved counting the mentions of research data and citations in the full text of article for each discipline.

- Objective: To examine data reuse practices in various STEM fields.

I gained direct insight into formal and informal data citation based on data reuse in STEM subjecting the full text of articles to semi-automated content analysis. The citation processes considered were formal and informal. Formal citations appeared in the references section of articles indexed by the WoS. The assumption was that sharers are more likely to receive scholarly credit for their data in the form of a recorded citation when references to appear in the references section. Informal citation refers to situations in which shared data is acknowledged outside of the references section such as in the acknowledgements section as a scholar's courtesy (Cronin, 1995; Cronin, 2001) or in passing in the main text. Citation databases such as the WoS do not index informal references in published articles, and formal citation has been found to be low in some disciplines. Thus, for example, in 2014, only around 12% of data citations in oceanography articles were formal, the other88% being informal (Belter, 2014).

By examining over 15,000 instances in 705 articles in which data sharing and reuse were cited, I determined that research data were widely cited informally in the main text, especially in the methodology section. This finding suggests that the total number of data mentions for data sharing and reuse, both formal and informal, should be taken into account in order to assess the impact of research data in scientific disciplines accurately.

Table 38 presents an overview of the instances of formal and informal data citation related to data reuse and sharing. Again, formal data citation of data sharing and reuse (9.6%) occurred less frequently than informal data citation (90.4%) in the STEM fields. Further, the frequency of data reuse (51.1%) was similar to that of data sharing (50.7%)- simultaneous -data reuse and sharing were not counted. Documentation of data reuse and sharing was fount most frequently in the main text (72.4%) of the associated literature of research data especially in the methods section of full text articles, followed by supplementary material (10.5%), and references (9.6%). Neither version numbers nor permanent identifiers such as DOIs, were observed in the formal data citations, nor was version information included as part of the titles of the articles.

Table 38 Location and practices for data reuse and sharing in STEM fields

| citation type | location in journals | data reuse | data reuse/ repeat | data reuse/ sharing | data sharing | data sharing/ repeat | total |
|---|---|---|---|---|---|---|---|
| informal data citation (90.4%) | abstract (0.9%) | 14 | 4 | 0 | 5 | 0 | 23 |
|  | acknowledgments (3.6%) | 56 | 17 | 0 | 20 | 2 | 95 |
|  | footnotes (3%) | 41 | 5 | 0 | 27 | 7 | 80 |
|  | main text (72.4%) | 1,002 | 394 | 15 | 407 | 112 | 1,930 |
|  | supplementary material (10.5%) | 45 | 12 | 0 | 196 | 27 | 280 |
| formal data citation (9.6%) | references (9.6%) | 204 | 31 | 0 | 20 | 1 | 256 |
| totals |  | 1362 | 463 | 15 | 675 | 149 | 2,664 |

Table 39 displays the comparative analysis of informal data citation vs. formal data citation in bibliographies in STEM fields. The rates were found to vary across disciplines, but informal data citation was, again, more common than formal data citation in STEM fields.

Table 39 Comparative analysis of informal data citation vs. formal data citation in bibliographies in STEM

| **discipline** | **informal data citation** (total data citations, percentage) | **formal data citation** (total data citations, percentage) |
|---|---:|---:|
| **astronomy/physics** | 338 (98%) | 9 (2%) |
| **biological sciences** | 1,342 (95.4%) | 65 (4.6%) |
| **chemistry** | 137 (93.2%) | 10 (6.8%) |
| **computing** | 90 (86.5%) | 14 (13.5%) |
| **earth sciences** | 228 (86.4%) | 36 (13.6%) |
| **engineering** | 45 (69.2%) | 20 (30.8%) |
| **mathematical sciences** | 115 (60.9%) | 52 (31.1%) |
| **technology** | 113 (69.3%) | 50 (30.7%) |

Tables 40 to 45 demonstrate formal and informal data citation by each discipline in detail. Viewing the data citation phenomenon independently, I observed disciplinary differences sharing and reuse across STEM disciplines. (Once again, simultaneous data reuse and sharing were not counted for any of the disciplines).

Table 40 summarizes these results for astronomy/physics. Informal data citation (98%) was 49 times more frequent than formal data citation (2%), while there was relatively little difference between the frequencies of data reuse (40.9%) and data sharing (46.4%). These latter practices were cited most frequently in the main text (81%), and occasionally in footnotes (7.2%) or supplementary material (4.3%).

Table 40 Location and practices for data reuse and sharing in astronomy/physics

| **citation type** | **location in publications** | **data reuse** | **data reuse/ repeat** | **data reuse/ sharing** | **data sharing** | **data sharing/ repeat** | **totals** |
|---|---|---|---|---|---|---|---|
| informal data citation (98%) | abstract (2%) | 5 | 1 | 0 | 1 | 0 | 7 |
| | acknowledgments (2.9%) | 8 | 1 | 0 | 1 | 0 | 10 |
| | footnotes (7.2%) | 14 | 1 | 0 | 10 | 0 | 25 |
| | main text (81%) | 113 | 33 | 1 | 129 | 5 | 281 |
| | supplementary material (4.3%) | 0 | 0 | 0 | 14 | 1 | 15 |
| formal data citation (2%) | references (2.6%) | 2 | 1 | 0 | 6 | 0 | 9 |
| **totals** | | 142 | 37 | 1 | 161 | 6 | 347 |

Table 41 summarizes these results for the biological sciences. Informal data citation (95.4%) was found to be around 20 times more frequent than formal (4.6%), consistent with the previous findings (Park, You, & Wolfram, 2018). Data reuse (47.8%) occurred more frequently than data sharing (22.1%), and both practices were cited most frequently in the main text (75.6%) and supplementary material (13.4%) or references (4.6%).

Table 41 Location and practices for data reuse and sharing in biological sciences

| **citation type** | **location in publications** | **data reuse** | **data reuse/ repeat** | **data reuse/ sharing** | **data sharing** | **data sharing/repeat** | **totals** |
|---|---|---|---|---|---|---|---|
| informal data citation (95.4%) | abstract (0.9%) | 7 | 3 | 0 | 3 | 0 | 13 |
| | acknowledgments (3.3%) | 24 | 9 | 0 | 11 | 2 | 46 |
| | footnotes (2.2%) | 20 | 4 | 0 | 5 | 2 | 31 |
| | main text (75.6%) | 529 | 274 | 8 | 169 | 84 | 1,064 |
| | supplementary material (13.4%) | 40 | 12 | 0 | 120 | 16 | 188 |
| formal data citation (4.6%) | references (4.6%) | 52 | 9 | 0 | 4 | 0 | 65 |
| **totals** | | 672 | 311 | 8 | 311 | 105 | 1,407 |

Table 42 summarizes the results for chemistry. Informal data citation (93.2%) occurred around 14 times more frequently than formal (6.8%), and data reuse (29.3%) occurred less frequently than data sharing (50.3%). The latter practices were cited most often in the main text (49%) and supplementary material (35.4%).

Table 42 Location and practices for data reuse and sharing in chemistry

| **citation type** | **location in publications** | **data reuse** | **data reuse/ repeat** | **data reuse/ sharing** | **data sharing** | **data sharing/ repeat** | **totals** |
|---|---|---|---|---|---|---|---|
| informal data citation (93.2%) | abstract (0%) | 0 | 0 | 0 | 0 | 0 | 0 |
| | acknowledgments (4.1%) | 3 | 0 | 0 | 3 | 0 | 6 |
| | footnotes (4.8%) | 0 | 0 | 0 | 5 | 2 | 7 |
| | main text (49%) | 32 | 6 | 1 | 22 | 11 | 72 |
| | supplementary material (35.4%) | 3 | 0 | 0 | 41 | 8 | 52 |
| formal data citation (6.8%) | references (6.8%) | 5 | 2 | 0 | 3 | 0 | 10 |
| **totals** | | 43 | 8 | 1 | 74 | 21 | 147 |

Table 43 summarizes the results for computing, in which informal data citation (86.5%) occurred around 4 times more frequently than formal (13.5%), and data reuse (68.3%) was more frequent than data sharing (19.2%) These latter practices were most often cited in the main text (75%) and occasionally in the references (13.5%) or supplementary material (9.6%).

Table 43 Location and practices for data reuse and sharing in computing

| citation type | location in publications | data reuse | data reuse/ repeat | data reuse/ sharing | data sharing | data sharing/ repeat | totals |
|---|---|---|---|---|---|---|---|
| informal data citation (86.5%) | abstract (0%) | 0 | 0 | 0 | 0 | 0 | 0 |
| | acknowledgments (1%) | 1 | 0 | 0 | 0 | 0 | 1 |
| | footnotes (1%) | 0 | 0 | 0 | 1 | 0 | 1 |
| | main text (75%) | 57 | 10 | 2 | 9 | 0 | 78 |
| | supplementary material (9.6%) | 0 | 0 | 0 | 10 | 0 | 10 |
| formal data citation (13.5%) | references (13.5%) | 13 | 1 | 0 | 0 | 0 | 14 |
| **totals** | | 71 | 11 | 2 | 20 | 0 | 104 |

Table 44 summarizes these results for the earth sciences, in which informal data citation (86.4%) occurred around 6 times more frequently than formal (13.6%), and data reuse (65.2%) was around 3 times more frequent than data sharing (21.2%). The latter practices were cited most often in the main text (71.6%) and occasionally in the references (13.6%) and acknowledgments (10.2%). Formal data citation, mostly related to data reuse, was relatively high in the earth sciences compared with other disciplines.

Table 44 Location and practices for data reuse and sharing in earth sciences

| citation type | location in publications | data reuse | data reuse/ repeat | data reuse/ sharing | data sharing | data sharing/ repeat | totals |
|---|---|---|---|---|---|---|---|
| informal data citation (86.4%) | abstract (0.4%) | 1 | 0 | 0 | 0 | 0 | 1 |
| | acknowledgments (10.2%) | 17 | 7 | 0 | 3 | 0 | 27 |
| | footnotes (1.9%) | 1 | 0 | 0 | 3 | 1 | 5 |
| | main text (71.6%) | 132 | 39 | 3 | 12 | 3 | 189 |
| | supplementary material (2.3%) | 1 | 0 | 0 | 5 | 0 | 6 |
| formal data citation (13.6%) | references (13.6%) | 20 | 11 | 0 | 5 | 0 | 36 |
| **totals** | | 172 | 57 | 3 | 28 | 4 | 264 |

Table 45 summarizes these results for engineering, in which informal data citation (69.2%) was more frequent than formal (30.8%) but relatively less frequent than most STEM fields. Data reuse (56.9%) was more than 3 times higher than data sharing (18.5%). in the latter practices were cited most often in the main text (64.6%), references (30.8%) and rarely in the footnotes (3.1%).

Table 45 Location and practices for data reuse and sharing in engineering

| **citation type** | **location in publications** | **data reuse** | **data reuse/ repeat** | **data reuse/ sharing** | **data sharing** | **data sharing/ repeat** | **totals** |
|---|---|---|---|---|---|---|---|
| informal data citation (69.2%) | abstract (0%) | 0 | 0 | 0 | 0 | 0 | 0 |
| | acknowledgments (1.5%) | 1 | 0 | 0 | 0 | 0 | 1 |
| | footnotes (3.1%) | 0 | 0 | 0 | 1 | 1 | 2 |
| | main text (64.6%) | 24 | 8 | 0 | 9 | 1 | 42 |
| | supplementary material (0%) | 0 | 0 | 0 | 0 | 0 | 0 |
| formal data citation (30.8%) | references (30.8%) | 12 | 6 | 0 | 2 | 0 | 20 |
| **totals** | | 37 | 14 | 0 | 12 | 2 | 65 |

Table 46 summarizes these results for the mathematical sciences, in which informal data citation (68.9%) was around twice as frequent as formal (31.1%) but relatively less frequent than in most STEM fields. Data reuse (78.4%) was around 5 times more frequent than data sharing (14.4%) The latter practices were cited most often in the main text (61.7%), occasionally in the references (31.1%), and rarely in the supplementary material (3%) or footnotes (2.4%).

Table 46 Location and practices for data reuse and sharing in mathematical sciences

| citation type | location in publications | data reuse | data reuse/ repeat | data reuse/ sharing | data sharing | data sharing/ repeat | totals |
|---|---|---|---|---|---|---|---|
| informal data citation (68.9%) | abstract (0%) | 0 | 0 | 0 | 0 | 0 | 0 |
| | acknowledgments (1.8%) | 2 | 0 | 0 | 1 | 0 | 3 |
| | footnotes (2.4%) | 4 | 0 | 0 | 0 | 0 | 4 |
| | main text (61.7%) | 72 | 6 | 0 | 21 | 4 | 103 |
| | supplementary material (3%) | 1 | 0 | 0 | 2 | 2 | 5 |
| formal data citation (31.1%) | references (31.1%) | 52 | 0 | 0 | 0 | 0 | 52 |
| **totals** | | 131 | 6 | 0 | 24 | 6 | 167 |

Table 47 summarizes these results for technology, in which informal data citation (69.3%) was around twice as frequent as formal (30.7%) but relatively less frequent than in most STEM fields. Data reuse (57.7%) was around twice as frequent as data sharing (27.6%). The latter practices were most often cited in the main text (62%) or references (30.7%) and rarely in the footnotes (3.1%)

Table 47 Location and practices for data reuse and sharing in technology

| **citation type** | **location in publications** | **data reuse** | **data reuse/ repeat** | **data reuse/ sharing** | **data sharing** | **data sharing/ repeat** | **totals** |
|---|---|---|---|---|---|---|---|
| informal data citation (69.3%) | abstract (1.2%) | 1 | 0 | 0 | 1 | 0 | 2 |
| | acknowledgments (0.6%) | 0 | 0 | 0 | 1 | 0 | 1 |
| | footnotes (3.1%) | 2 | 0 | 0 | 2 | 1 | 5 |
| | main text (62%) | 43 | 18 | 0 | 36 | 4 | 101 |
| | supplementary material (2.5%) | 0 | 0 | 0 | 4 | 0 | 4 |
| formal data citation (30.7%) | references (30.7%) | 48 | 1 | 0 | 1 | 0 | 50 |
| **totals** | | 94 | 19 | 0 | 45 | 5 | 163 |

To summarize the results relating to RQ4:

- Formal data citation of data sharing and reuse at the bibliographic-level (i.e., articles in the WoS) accounted for less than 10% of all data citations, the vast majority being informal.
- Data citation practices were diverse across STEM disciplines.
- The frequency of informal compared with formal data citation was very high in astronomy/physics (98%), the biological sciences (95.4%), chemistry (93.2%), computing (86.5%), and the earth sciences (86.4%). By contrast, engineering (69.2%), the mathematical sciences (60.9%), and technology (69.3%) had relatively low levels of informal data citation.
- The main text was the most common location for the documentation of data sharing and reuse.
- The frequency of data sharing and reuse practices in the main text varied across disciplines, from relatively high in astronomy/physics (81%), the biological sciences (75.6%), earth sciences (71.6%), engineering (64.4%), and technology (61.7%) to relatively low in chemistry (49%).

### 4.2.5. RQ5: To what extent do STEM disciplines support interdisciplinary data citation?

Table 48 displays the journals for pairs of STEM disciplines with the potential for shared knowledge discovery and scientific measurement in the future. Examination of the journals (i.e., formal publication venues) serving the 8 disciplines that were the focus of this study revealed other aspects of their relations. Thus, for example, the 589 journals in the biological sciences displayed a relatively high degree of interdisciplinarity based on the number of fields cited in their articles.

Research in STEM, as discussed in detail earlier, is no longer limited to a single field, but is rather often interdisciplinary nature. As just mentioned, and as Table 48 indicates, the biological sciences received the most citations among the eight fields – and is noteworthy that this result reflects the disciplinary breadth of these citations as well. The results also indicated that researchers in some STEM disciplines tended not to collaborate with those in other disciplines. Thus, no interdisciplinarity was observed for engineering and relatively little for chemistry or computing. The phenomenon was not observed in the dedicated journals. The average number of cited subject categories per journal varied depending on the category.

Table 48 Total number of citations from other fields in STEM

| **field** | **sum of records** |
|---|---|
| astronomy/physics | 53 |
| biological sciences | 6,014 |
| chemistry | 30 |
| computing | 1 |
| earth sciences | 591 |
| engineering | 0 |
| mathematical sciences | 279 |
| Technology | 54 |
| **grand total** | **7,007** |

Table 49 displays the total number of citations of astronomy/physics by other fields. This discipline as constructed in this study is by nature interdisciplinary, being the integration of astronomy and physics. In astronomy/physics, then, 53 citations in 18 articles fell into 7 subject categories. Researchers in astronomy/physics cited sources within and beyond their own subject areas, in the latter case mainly ones in neighboring rather than distant disciplines.

Table 49 Total number of citations from other fields in astronomy/physics

| **field** | **sum of records** |
|---|---|
| biology & biochemistry | 1 |
| chemistry | 11 |
| engineering | 2 |
| pharmacology & toxicology | 17 |
| Physics | 2 |
| plant & animal science | 7 |
| space science | 13 |
| **totals** | **53** |

Table 50 indicates that the biological sciences received the most citations from other fields in other STEM fields, meaning that researchers in this field engaged particularly actively in interactions with these other disciplines. Research data tended to be more cited by other papers from diverse disciplines in biological sciences. Specifically, research data in the biological sciences were cited by 6,014 articles across 19 subject categories. This indicated that researchers this discipline had access to the greatest diversity of opinions and thinking among the eight STEM disciplines studied here.

Table 50 Total number of citations from other fields in biological sciences

| **field** | **sum of records** |
|---|---|
| agricultural sciences | 29 |
| biology & biochemistry | 508 |
| chemistry | 1 |
| clinical medicine | 2,365 |
| computer science | 2 |
| engineering | 5 |
| environment/ecology | 50 |
| geosciences | 24 |
| Immunology | 502 |
| materials science | 3 |
| mathematics | 4 |
| microbiology | 468 |
| molecular biology & genetics | 570 |
| multidisciplinary | 509 |
| neuroscience & behavior | 204 |
| pharmacology & toxicology | 211 |
| plant & animal science | 165 |
| psychiatry/psychology | 60 |
| social sciences, general | 334 |
| **totals** | **6,014** |

Chemistry did not receive citations from other fields. Table 51 provides evidence that other disciplines did not cite chemistry-based data. The small number of recorded citations indicates that the chemistry data were largely siloed.

Table 51 Total number of citations from other fields in chemistry

| **field** | **sum of records** |
|---|---|
| chemistry | 30 |
| **totals** | **30** |

Table 52 displays the total times of citation computing received by other fields. Data citation practice in computing was rare and only from one other discipline. Research data were cited by only one article in physics in the WoS.

Table 52 Total number of citations from other fields in computing

| **field** | **sum of records** |
|---|---|
| physics | 1 |
| **totals** | **1** |

The findings presented in Table 53 indicates that researchers in other fields cited work in the earth sciences, both those from neighboring disciplines (e.g., agricultural sciences and geosciences) and more distant ones (e.g., psychiatry/psychology and the social sciences). This finding deserves further exploration -variations in citation practices over time and owing to differences in gender, age, and geographical location were beyond the scope of this study (discussed below).

Citations of data from earth sciences research were quite diverse, occurring in 591 articles in 16 subject categories in the WoS. Most of these data citation were from articles in geosciences journals (200 articles), followed by multidisciplinary (188 articles) and environment/ecology journals (81 articles).

Table 53 Total times cited counts by other fields in earth sciences

| **field** | **sum of records** |
|---|---|
| agricultural sciences | 10 |
| biology & biochemistry | 2 |
| chemistry | 5 |
| clinical medicine | 9 |
| engineering | 8 |
| environment/ecology | 81 |
| geosciences | 200 |
| immunology | 11 |
| microbiology | 5 |
| molecular biology & genetics | 10 |
| multidisciplinary | 188 |
| physics | 1 |
| plant & animal science | 42 |
| psychiatry/psychology | 6 |
| social sciences, general | 10 |
| space science | 3 |

| | |
|---|---|
| **totals** | **591** |

The findings presented in Table 54 indicate that research data in engineering were not cited by any articles in other subject categories in the WoS. This means that engineering was siloed and that data citation was not being practiced in engineering.

Table 54 Total times cited counts by other fields in engineering

| **field** | **sum of records** |
|---|---|
| n/a | n/a |
| **totals** | **0** |

Table 55 displays the citation of data from research in the mathematical sciences cited by researchers in other fields. In fact, these other fields were numerous with the 279 citations representing 11 subject categories in the WoS. Data was mostly cited by articles in multidisciplinary (175 articles), followed by geosciences (28 articles) and environmental/ecology journals (15 articles).

Table 55 Total times cited counts by other fields in mathematical sciences

| **field** | **sum of records** |
|---|---|
| biology & biochemistry | 2 |
| clinical medicine | 9 |
| environment/ecology | 15 |
| geosciences | 28 |

| | |
|---|---|
| immunology | 11 |
| microbiology | 5 |
| molecular biology & genetics | 10 |
| multidisciplinary | 175 |
| plant & animal science | 10 |
| psychiatry/psychology | 6 |
| social sciences, general | 8 |
| **totals** | **279** |

Table 56 displays the citation counts connecting other fields to technology, which amounted to 54 articles across 13 subject categories in the WoS. Research data were mainly cited in articles in clinical medicine (12 articles), multidisciplinary (12 articles), and plant & animal sciences journals (9 articles).

Table 56 Total times cited counts by other fields in technology

| **field** | **sum of records** |
|---|---|
| agricultural sciences | 1 |
| biology & biochemistry | 2 |
| chemistry | 1 |
| clinical medicine | 12 |
| engineering | 2 |
| geosciences | 5 |
| materials science | 1 |

| | |
|---|---|
| microbiology | 1 |
| molecular biology & genetics | 4 |
| multidisciplinary | 12 |
| pharmacology & toxicology | 1 |
| physics | 3 |
| plant & animal science | 9 |
| **totals** | **54** |

Table 57 displays the diversity and interdisciplinarity across STEM fields as an indicator of balance with respect to interdisciplinary data citation. This analysis was conducted to measure interdisciplinarity comparatively. In this study, variety refers to the number of cited fields in the WoS and balance to the distribution of citations among fields in the WoS. I applied the ESI fields in order to measure interdisciplinarity based on the diversity of an article's cited literature as indicated by the variety of ESI fields of the citing journals. I applied the Gini-index in order to measure balance (Nijssen, Rousseau, & van Hecke, 1998). However, as discussed by Leydesdorff (2018), the Gini-index does not measure variety, so I applied Leydesdorff's formula in order to measure interdisciplinarity using one formula in terms of the three aspects of diversity: (again, variety, balance and disparity).

Based on the citations received from other fields (i.e., interdisciplinarity based on the diversity of an article's cited literature that focuses on the variety of journals cited), the biological sciences received the most citations from other subject categories, followed by the earth sciences and technology. In order to measure inequality among STEM fields regarding interdisciplinary data citation, I applied Gini's diversity index as a measure of concentration (i.e., of inequality or balance), with 0 indicating complete equality and 1 complete inequality. Chemistry, computing and engineering showed complete equality (i.e., no concentration, meaning that the citations are equally distributed across journals). These three disciplines were then excluded from the measurement of diversity because they did not attract citations from more than one field.  Based on the disciplinary diversity in the references (i.e., the Gini-index), the mathematical sciences showed the most inequality (i.e., the most diversity), followed by astronomy/physics, the earth sciences, the biological sciences and technology, meaning that citations were not equally distributed across journals. For the Leydesdoff interdisciplinarity formula outcomes, the earth sciences showed the highest level of interdisciplinarity, though more fields cited work in the biological sciences.

Table 57 Diversity and interdisciplinarity in STEM

| **Discipline** | **number of ESI fields** | **Gini-index** | **Leydesdorff's interdisciplinarity calculation** |
|---|---|---|---|
| astronomy/physics | 7 | 0.8173 | 0.1821 |
| biological sciences | 19 | 0.7112 | 0.1695 |
| chemistry | 1 | 1 | 0 |
| computing | 1 | 1 | 0 |
| earth sciences | 16 | 0.7852 | 0.2728 |

| | | | |
|---|---|---|---|
| engineering | 0 | 1 | 0 |
| mathematical sciences | 11 | 0.8229 | 0.0885 |
| technology | 13 | 0.6987 | 0.1826 |

These findings for RQ5, which relates to disciplinary knowledge exchange, provide support for a growing trend in interdisciplinary diffusion across research data, journal articles and research areas.

To summarize:

- Interdisciplinarity in data citation varied across STEM disciplines.
- The biological sciences received the most citations from other disciplines, with research data having been cited by 6,014 journals across 19 subject categories in the WoS. Engineering, by contrast, showed no interdisciplinarity and chemistry and computing very little.
- The earth sciences showed the greatest degree of interdisciplinarity based on Leydesdorff's formula. Although it had fewer citing disciplines than the biological sciences, those citations were more evenly distributed.

# Chapter 5 DISCUSSION

In this chapter, I discussed the main findings of this study. In doing so, I present answers to the five research questions regarding variation in data sharing and reuse practices and their impact on data citation varied across STEM fields (i.e., domain-specific or discipline-specific practices) and interdisciplinary data citation. The contributions made by and limitations of this study are also addressed in this chapter.

Specifically, I discuss in turn - (1) data sharing, (2) data type, (3) data reuse, (4) author self-citation and recitation, (5) disciplinary differences, and (6) interdisciplinarity.

## 5.1. The impact of data sharing

Data sharing has become increasingly common in the STEM fields in recent years. It was most frequent in the biological sciences, possibly owing to the relatively early adoption of data sharing requirements by the NIH (in 2003) compared with influential organizations in other fields (e.g., the NSF adopted such requirements in 2011). This finding is consistent with that of Piwowar and Chapman (2010), who observed more frequent data sharing in the biomedical fields than in others, again seemingly in connection with the implementation of the NIH's data-sharing mandate.

The finding that the frequency of data sharing varies across STEM disciplines raises the question of how the practice must be promoted. One approach would be to provide formal credit to data sharers that could be adduced as evidence of scholarly activity in professional contexts, such as consideration for tenure and promotion. The advisability of such an approach receives further support from the observation by Andreoli-Versbach and Mueller-Langer (2014) of both an increase in the citation of articles published in journals that had mandatory data-sharing policies and changes in personal attitudes toward the open science movement, which promotes data sharing.

Another finding was that data reuse was five times more frequent than data sharing in the mathematical sciences, while in chemistry the situation was reversed, with data sharing occurring twice as often as data reuse. A possible explanation for this finding is that perceptions of the importance of data sharing vary across disciplines. Thus, in one recent study, researchers in chemistry considered data sharing a crucial factor in novel scientific findings, while researchers in the mathematical sciences did not (Kim, 2013).

It was further observed in the present study that each discipline relied on a few discipline-specific repositories for data sharing, as indexed by the DCI. Such third-party digital repositories were preferred over the websites of journal publishers' or individuals. For data sharers, the choice of a leading repository in which to preserve their research data and to receive formal scholarly credit can be a real concern and merits further investigation. One advantage of discipline-specific repositories is that users can quickly narrow their searches through tailored matches to data within that discipline. Researchers must, therefore, familiarize themselves with the controlled vocabularies and subject categories used in particular repositories.

.

## 5.2. The impact of author self-citation and recitation

To the best of my knowledge, this is the first attempt to examine author self-citation and recitation in the context of data citation in the STEM fields. Author self-citation is the situation in which authors cite their own previous work in subsequent articles. In terms of specific statistics, the analysis showed that the average rates of author self-citation and recitation were similarly low at the data-level (3.94%) and article-level (3.88%). This result could indicate that few researchers were reusing research data in general, or that they were reusing shared data other than their own. It is also important to keep in mind that only recently have journals begun to implement data sharing policies, *PLoS ONE*' in 2014 (Silva, 2014) and *Nature Research* (2017), *Science* (2017), and *Elsevier* (2017) following suit in 2017. The time required to adjust to these policies may in part explain the low rate of author self-citation and recitation (an average frequency of 3.91% for both practices.

## 5.3. The impact of data type

With one exception, all of the data types identified in this study were quantitative in nature. This finding is attributable to the fact that qualitative data tend to be viewed with skepticism in the STEM fields while being more accepted in the social sciences (Mason, 2007; Mauthner & Parry, 2009; Yoon, 2014). There may also be ethical considerations regarding personally identifiable data in some fields, such as biomedicine. The one piece of qualitative data identified here was an interview transcript. When such documents are shared, direct identifiers—names, email addresses, date of birth, addresses, and so on—must be removed in order to preserve the privacy of the participants. Indirect identifiers, which make it possible to identify individual participants or patients by crossing the data with other datasets, must also be removed. Thus, for instance, the sharing of interview transcriptions might require reading through many pages of text from multiple participants to see whether they mention names or key dates, the latter of which could serve as indirect identifiers, as in the case of admittance and delivery dates in hospitals. Other considerations related to the sharing of qualitative data include confidentiality and disclosure risks for research involving minors, these again being issues that tend to arise in the context of the social sciences. Jeng (2017) has noted such other challenges as the low awareness, the time and effort required for data preservation, and the confidentiality concerns and confidence of individual primary investigators. These latter considerations and challenges pertain to both the social sciences and STEM fields; and indeed the open science movement encourages data sharing across all disciplines. In any cases, the conflicting ethical considerations of confidentiality and open science can complicate data reuse practices

Redundancies in the classification of data types were also observed. Thus, for instance, the data type “dataset” was also categorized as “data/dataset”, so that  the classification scheme used in Data-Planet caused classification redundancies between two levels regarding the resource types in DataCite.com specifically the resource type “datatype” was also found as the subtype “dataset” in the latter repository owing to inconsistencies in the two classified records based on the repository’s scheme. Further, some repositories, such as E-Periodica and ETH E-Collection, have only “text” records (Robinson-Garcia, Mongeon, Jeng, & Costas, 2017). Although such inconsistencies are beyond the scope of this study, they deserve further study.

The most-cited data types varied depending on the discipline (Table 13). The greatest variety in this respect was observed in the biological sciences and the earth sciences and the least in engineering and the mathematical sciences. The widespread use of multiple data types by various researchers across diverse disciplines has made the assessment of the integrity and trustworthiness of research data extremely complicated when it comes to documenting, tracking and maintaining the data in a single workflow (Darch et al., 2015).

Accurate classification of data types is, then, a crucial part of facilitating data sharing and reuse and, therefore, of scientific reproducibility. Classifying data types can, however, be time-consuming for data sharers because the process requires prior knowledge of the data types. What appears to be needed in order to avoid imposing an additional workload on data sharers is an automatic (i.e., machine-actionable) identification system that uses a standardized data type across disciplines. Imposing this level of uniformity would not be easy. To begin with machine-actionable classification schemes obviously require machine-readable definitions and datasets that include dynamic data can have multiple structures. Retrieving the needed data types from a system (e.g., a federated system) for data reusers would be especially challenging in the absence of precision with regard to the data types shared. Interfaces and technologies that span multiple disciplines would need to be developed for the reanalysis of various data types. Data visualization in a federated system could, however, serve as a tool for reanalysis by accessing the various data types in distributed repositories that, for instance, rely on cloud storage. Additionally, more detailed steps regarding interoperability across cyberinfrastructures and shared technologies of a federated system need to be considered on an ongoing basis.

## 5.4. The impact of data reuse

A further finding is that formal citation was not common in scholarly publications in the STEM fields when data were reused (See Table 38). At the disciplinary community level, organizations often lack standard data citation guidelines and the guidelines that are in place differ from one organization to the next. Even when data are cited, the citation rarely provides accurate access information linked to actual data. In the assessment of formal data citation, the heterogeneous and unstructured nature of research data suggests few straightforward solutions. Style manuals, such as the one published by the American Psychological Association (APA) and the Chicago Manual of Style, though, do provide guidelines for the formal citation of datasets, the APA's having done so since 1983(American Psychological Association, 1983). In this study, few instances of data citations were found to follow such style guidelines by providing permanent identifiers such as DOIs. This finding indicates a lack of awareness on the part of researchers that data are as much a citable source as more traditional material, such as articles. The guidelines for research data provided by the style manuals evolve in step with publishers' requirements for the formatting citations – a situation that highlights the importance of assigning a DOI to research data with its citing articles (i.e., the landing pages for the research data). Among the challenges of using DOIs as permanent identifiers is that the fact that not all publishers turn the DOI-prefixed form into a hyperlink, for which reason Hourclé, Chang, Linares and Palanisamy (2012) recommended using the HTTP URL form (http://dx.doi.org/10...).

Informal data citation was found to be more common than formal data citation in the STEM fields (Table 38). It sometimes occurred in the acknowledgments section of articles, which is where, according to Cronin (1995; 2001), authors express and discuss norms, patterns, and trends. In any case, the challenge of the large amount of labor required to collect data from unstructured text in published literature. In sum, the acknowledgments section is a poor choice as a place to cite data because of inconsistent formatting and other practical difficulties.

Time gaps for data reuse impact need to be considered. Previous studies have reported that the increased citation rate in core astrophysical journals such as *Astrophysical Journal* published in 2010 articles are linked to research data (Drachen, Elleggard, Larsen, & Fabricius, 2016). Whether the finding here that the lowest rate of formal data citation occurred in astronomy/physics (2.59%) is due to the time gaps (i.e., a 5-year time span by Drachen et al. vs. 15-year time span in the present study) is again beyond the scope of this dissertation and left for future study.

Attention also needs to be given to measurement of the reuse of subsets of data. A possible approach would be to use the precise version and time stamp of a dataset as its permanent identifier.

By way of precedent, Katz and Smith (2015) proposed, as a means to give partial credit to indirect contributors in the context of software, a form of transitive credit that would involve assigning varying amounts of credit for both research software and contributors using machine-readable JavaScript Object Notation- Linked Data (JSON-LD).

The citation of data reuse at the data level (i.e., the establishment of usage metrics for research data) becomes problematic when the data reside behind institutional or corporate firewalls. Thus, for instance, corporations (i.e., reusers) may download open data from a repository and then reuse it while it is stored on their own in-house systems. In the absence of an active script in the open data itself that allows for the counting of every single reuse, even the application of offline reuse metrics poses a challenge. This challenge might be surmountable through the use of event-based data usage metrics that standardizes the ways in which downloads and views are counted (Fenner et al., 2018) and including usage statistics that keep track of access events (DataCite, 2018, Data Observation Network for Earth (2018). Whether a partial download (e.g., involving data breakage owing to massive volumes of data) or live streaming of data should be weighted equally as data usage for reuse metrics is a related issue deserving consideration.

## 5.5. The impact of disciplinary differences

Data sharing and reuse practices and tendencies were found to be largely field-specific, with particular skewedness in the biological sciences. The prevalence of data sharing in genomics (which was classified among the biological sciences in this study), for instance, is a well-established phenomenon attributable to the relatively early development of the necessary infrastructure in this field (Anagnostou et al., 2015; Choudhury, Fishman, McGowan, & Juengst, 2014; Kaye, Heeney, Hawkins, de Vries, & Boddington, 2009; Mongeon, Robinson-García, Jeng, & Costas, 2017). The relatively high rate of data sharing in the biological sciences and the disciplinary unevenness of sharing across STEM fields should be interpreted with care, however, for the extensive use of proprietary or sensitive data can complicate data sharing in certain disciplines, such as medicine, as mentioned earlier. In any case, the variety of data types needs to be considered when accounting for disciplinary differences in data sharing.

## 5.6. The impact of interdisciplinarity

Scholars have shown increased interest in understanding the mechanisms that facilitate knowledge transfer across disciplines. The study of interdisciplinarity and knowledge diffusion in scholarly communication helps to clarify factors that contribute to gains and losses in knowledge over time and within and across disciplines. However, there has been almost no research into interdisciplinary data citation to date, though the impact of interdisciplinarity on research data has studied in terms of the role of article-level citation networks on the diffusion of research data across all subject categories in the WoS.

In this study, the biological sciences received the most citations from the most fields among the STEM disciplines analyzed (see Table 50). To be specific, roughly 86% of research data citations in the biological sciences came from articles published in other fields in the WoS. This result indicates that an enormous amount of research data used in the biological sciences was also used in research published in journals serving other disciplines. Another question thus arises that merits further study regarding whether certain disciplines tend in general to produce more citations than others.

Both computing (with a single citation from another discipline) and engineering (with no recorded data citations at all) were revealed in this study to be siloed disciplines when it comes to data citation. This finding may be artifact of the indexing practices of WoS. If not, it appears that researchers in these fields were largely indifferent to or unaware of the availably of shared data. In this respect, neither computing nor engineering had an impact on published research in other disciplines.

As seen in Table 53, this study revealed that practices of citations in the earth sciences occurred both in neighboring disciplines (e.g., in agricultural sciences and geosciences) and more distant ones (e.g., psychiatry/psychology and the social sciences). A previous study found, by contrast, that data in the earth sciences were cited primarily by journals in the physical sciences and multidisciplinary fields in Google Scholar (Chao, 2011), although both set of findings are consistent regarding the citations of earth sciences data in multidisciplinary journals. This result may be attributable to the fact that the other study addressed only publications included in NASA's Global Change Master Directory (GCMD; https://gcmd.nasa.gov/), while this one examined all published and indexed journals in the WoS, a much broader range of material. Another study found that datasets in oceanography (which was treated as part of the earth sciences in this study) were highly cited by researchers, in naturally, oceanography but also those in the atmospheric sciences, geosciences, and multidisciplinary fields, also as indexed in the WoS (Belter, 2014); this finding is also consistent with the findings presented here.

## 5.7. Limitations

Though I sought in designing this study to mitigate every possible limitation, I nevertheless recognize that certain constraints need to be taken into consideration when evaluating its significance. To begin with, I focused on STEM fields, so the findings presented here are not necessarily relevant to citation practices in for example, the social sciences or humanities. In any case, it is hoped that this exploratory study has shed light on the tendencies of researchers in various STEM fields regarding the sharing, , reuse, and citation of research data.

Owing to citation delays, it may be difficult to capture certain relationships among publications at the levels of data, article and discipline. There may also be concerns relating to the dependence of repositories on reports of findings.

A further potential limitation of this study concerns the indexing feature of the DCI used here to identify formal citations. The DCI allows users to download a maximum of 100,000 records per discipline (as does the WoS), a number that may be insufficient to gather a representative sample. The focus here on STEM fields, in which such citation is most prevalent, was however, deemed reasonable given that data citation has only recently begun to be investigated.

There is also some reason for concern with respect to potential bias in the ESI's journal categories owing to its predefined category structure or taxonomy and to lack of consensus about the accuracy of the categorization systems used by particular journals (Wagner et al., 2011). As noted, all of the publications examined here were obtained from the WoS databases, so conference proceedings and papers were not included. Bias may likewise have resulted from the sampling rationale in terms of yielding a polarized sample given that I examined the work of only 30 prolific authors in each discipline; a significantly larger sample size would at least have allowed for a more robust analysis.

Lastly, characteristics other than the data sharing requirements of funding agencies and publishers were not considered in this study. These unexamined characteristics include the sectors represented by the various funding sources (e.g., governmental agencies, private companies, or individuals), the ownership of research data (in the context of which publishers' focus on searchable platforms may conflict with funders' desire to mandate management practices throughout the data life cycle), the age of shared data (thus the amount of time that elapses after data have been shared can influence digitization, loss, or changes in contact information), and technical obstacles to data sharing.

## 5.8. Implications

Despite these limitations, this exploratory study has helped to clarify the ways in which STEM researchers share, reuse, and cite research data and, most importantly, has captured the impact of data sharing and reuse on data citation in the STEM fields. Thus, it was revealed that the current reward system in STEM does not adequately recognize researchers' data sharing and reuse, a fact with significant implications for research and practice.

### 5.8.1. Practical implications

Practical implications of the findings presented here for researchers, decision-makers, funding agencies, and publishers include the importance of providing incentives to data sharers, promoting the sharing and reuse of research data, and understanding the needs of researchers. This study can be of particular use in this regard because it takes into account the distinct characteristics of various disciplines.

Much more work is needed to define best practices for research data sharing and reuse in the STEM fields, but it is clear enough that the awarding of formal scholarly credit varied greatly across disciplines. The insights offered here can thus be used to shape citation guidelines for individual disciplines, again so as to help the various stakeholder—for example, decision makers at university repositories—to identify citable research data.

This study also provides insights that should be of interest to project teams or companies that are currently developing data-level metrics databases regarding formal and informal data citation. Editors should be required to ensure that formal scholarly credit is given in the references section of articles. Further, based on the finding that formal data citation was twice as prevalent as informal data citation in engineering, the mathematical sciences, and technology, these disciplines should be consulted when journals formulate new policies. Thus, for instance, this study identified major repositories for each discipline in STEM fields (Table 16 and Table 17)k based on this finding, those who craft data policies for journals would be advised to include in them a list of suggested or recommended repositories. Authors who shared research data associated with their published articles in these major repositories can improve data citation opportunities by making their shared data visible and accessible.

When it comes to incentivizing researchers to share data, institutions should consider data citations in tenure review or promotion decisions, since the sharing of data serves the community and advances scholarly research. Researchers who share their research data, for their part, need to make clear how they are to be cited. They must also, again, take care to remove any information in shared research data - especially qualitative data - that could be used to identify, for example, medical patients without their consent.

### 5.8.2. Methodological implications

As discussed, the mixed methods approach (combining quantitative and qualitative methods) used here represents a contribution to the development of a theoretical basis for the field of information studies. The methodological implications of this study also include the elaboration of a semi-automatic text searching technique, the use of the Kruskal-Wallis test to account for the different numbers in the co-author self-citation and recitation groups, and the measurement of interdisciplinary data citation using Leydesdorff's calculation and the Gini-index.

Taking a moment to examine these issues in greater detail, first, the combination of quantitative and qualitative approaches allowed for deep insight into the impact of data sharing and reuse on data citation across multiple disciplines. The combination of automatic text extraction with human assessment using indicating terms represents a response to the ongoing challenges of data citation with associated articles, in particular the need to minimize manual assessment of the full text. Though this method does not identify all informal citations in the associated full text, it can significantly accelerate their discovery, and it also identifies more general terms for use, since newly identified terms tend not to include discipline-specific jargon. The Kruskal-Wallis test, by incorporating the group and individual levels, served to reveal key perspectives on co-author self-citation and recitation; no other study, to the best of my knowledge, has investigated both of these levels. Nor has another study applied the Gini-index and Leydesdorff's calculation to the measurement of interdisciplinary data citation by combining disparity, variety, and balance into a single formula.

# Chapter 6 CONCLUSION

In this chapter, I summarize the findings and suggest directions for future study. This dissertation has examined the impact of data sharing and reuse on data citation in the STEM fields, issues that have been under-investigated in the literature. The results shed light on the future development of data citation and stand to improve understanding of the sharing of research data in scholarly communication, with particular attention to the impact of data sharing and reuse on data citation in the STEM fields in terms of data type, discipline, and self-citation. The five research questions, which were introduced in Chapter 1 (Section 1.3), are reproduced here for the sake of convenience and completeness:

- RQ1: How prevalent is data sharing in various STEM disciplines as measured by formal data citation?
- RQ2: What types of STEM research data are formally cited most often?
- RQ3: How do author self-citation/recitation practices differ across STEM disciplines?)?
- RQ4: How do data reuse practices differ across STEM disciplines?
- RQ5: To what extent do the various STEM disciplines support interdisciplinary data citation?

## 6.1. Summary of the study

Each STEM discipline was found to have distinctive data sharing practices. Funding agencies played a major role in promoting data sharing in various STEM fields; thus, I observed marked increases in the frequency of data sharing after the NIH began requiring it. Data repositories likewise varied across disciplines, with the biological sciences tending to rely on governmental repositories, the earth sciences on discipline-specific repositories, and the mathematical sciences on discipline-independent repositories.

STEM researchers employed a wide array of data types; I documented some 454 in current use in the DCI. Quantitative research data were shared far more often than qualitative research data; in fact, only a single example of the latter, an interview transcript, was identified. Various data types were dominant across the individual STEM disciplines. Overall, the 10 most cited types in the DCI were data files, protein sequence data, crystallographic data, “blank” (that is, no specific data type provided by the record), software code, mass spectral data, crystal structure, molecular structure, Sequence Read Archive, filesets, and nuclear magnetic resonance. By discipline, the three most common types were, for astronomy/physics, mass spectral data, NMR results, and spectral data; for the biological sciences, RNA, protein sequence data, and SRA; for chemistry, crystal structure, crystallographic data, and molecular structure; for computing, software, code, and models; for the earth sciences, datasets, interactive resources, and GPS data; for engineering, test data, datasets, and GIS vector data; for the mathematical sciences, software, Matrix, and GEOID undulation on a grid; and for technology, datasets, filesets, and TIFF images.

Regarding author self-citation and recitation, a slight difference was found (0.06%), with a frequency of 3.94% for data-level and 3.88% for bibliographic-level citations. The average author self-citation and recitation frequency was 3.91%. The differences between the rates were greatest in computing (1.68%) and least in engineering (0.03%). At the data-level, author self-citation and recitation was highest in chemistry (7.29%) and lowest in computing (0.42%). At the bibliographic-level, author self-citation and recitation were most frequent in chemistry (6.92%) and least frequent in technology (1.38%).

Looking at the associations for author self-citation and recitation, bibliographic-level self-citations showed associations across various disciplines but, data-level self-citations did not. Some disciplines had no associations for self-citation. Disciplines that did not show associations with significant differences between the author self-citation and recitation rates were computing – earth sciences, computing – astronomy/physics, computing – biological sciences, computing – engineering and computing – chemistry, computing – engineering, technology – chemistry and mathematical sciences – chemistry at both the data and article levels. A difference was found in computing-engineering at the article level but not at the data-level.

It is one of the major findings of this dissertation that informal data citation was more common than formal data citation in the STEM fields and that the rates of both varied across disciplines. Informal data citation was most prevalent in astronomy/physics, followed by the biological sciences and chemistry. Specifically, the informal citation rates were 97.41% in astronomy, 95.38% in the biological sciences, 93.2% in chemistry, 86.36% in computing, 60.86% in the mathematical sciences, and 69.33% in technology. When the full text contents of articles were examined, actual data reuse (51.1 % across all STEM fields) was similar for sharing (50.7%) at the bibliographic level (i.e., articles). It should again be observed that simultaneous data reuse and sharing was not counted. To summarize the results by discipline:

- data reuse (40.9%) was slightly more frequent than data sharing (46.4%) in astronomy/physics;
- data reuse (47.8%) was around twice frequent as data sharing (22.1%) in the biological sciences;
- data reuse (29.3%) was less frequent than data sharing (50.3%) in chemistry;

- data reuse (68.3%) was more than four times more frequent than data sharing (19.2%) in computing;
- data reuse (65.2%) was more than three times more frequent than data sharing (21.2%) in the earth sciences;
- data reuse (56.9%) was around three times more frequent than data sharing (18.5%) in engineering;
- data reuse (78.4%) was around five times more frequent than data sharing (14.4%) in the mathematical sciences; and
- data reuse (57.7%) was around twice as frequent as than data sharing (27.6%) in technology.

Disciplinary unevenness in data sharing, then, was found across STEM. This result had been biased for the awareness and demand for software in recent years.

The aim of this dissertation was to shed the light on research practices in the STEM fields from the perspective of actual data sharing and reuse. The results suggest certain strategies for identifying the best practices for data citation, sharing, and reuse and have implications for data citation guidelines and policies in the STEM fields and beyond. The findings should therefore be of interest to researchers, publishers, funding agencies, and research organizations.

## 6.2. Directions for future research

Data citation has only recently begun to be studied from the perspectives of data sharing and reuse, so numerous approaches to this phenomenon have yet to be explored. Taking into account also the limitations to this study discussed in the previous chapter, the following avenues for future research appear to be particularly fruitful.

To begin with, the diffusion of specific geographic allocations could be explored as a means to capture research activities globally. Examining research activities at the geographic level in a detailed and timely manner could help to elucidate the knowledge diffusion process in general and user behaviors with respect to data citation in particular. Also, informative would be a longitudinal study reproducing and extending the findings presented here by tracking dynamic knowledge diffusion through data citation. Because data citation practices relating to data sharing and reuse may be more widespread and frequent in the future, the diversity in terms of the prevalence of data sharing as measured by data citation across the STEM fields deserves scrutiny, particularly with regard to the use of proprietary or sensitive data (Mongeon, Robinson-García, Jeng, & Costas, 2017). Further attention to these issues is also needed because the scope of this study did not allow for examination of the impact of the number of co-authors and self-citations on data citation.

Turning now to publication type, over 90% of data sharing as measured by citation occurred through journals rather than conference proceedings or books (Park & Wolfram, 2017). This result deserves careful consideration, for conference proceedings have generally been regarded as the primary venue for the dissemination of scholarship and research in such rapidly advancing areas as computer science. A possible explanation for the result found here is the adoption of more or less strict data sharing policies by high impact journals, data and otherwise.

Regarding the individual disciplines, the finding that data sharing was most prevalent in the biological sciences also invites further study. This finding may be attributable either to the DCI's indexing feature or to the gradual adoption of data sharing requirements by the major funding agencies—once more, the NIH issued this mandate in 2003 and the NSF eight years later. Future research could also extend the scope of the inquiry beyond STEM, that is, to the social sciences and humanities, again with an eye to similarities and differences across disciplines. One major difference in this respect is already apparent, namely the difficulty of sharing and reusing the qualitative data that are central to much non-STEM research (Yoon, 2014), as has indeed been demonstrated for studies based on interviews or ethnography (Faniel & Jacobsen, 2010; Wallis, Rolando, & Borgman, 2013). In any case, further research remains to be done on data sharing, reuse, and citation within and across the STEM fields in order to build on this exploratory study.

## 6.3. Final comments

In this era of open science, the wider availability, accessibility and reusability of open data are fundamental to and crucial for efficient scholarly communication and therefore for scientific progress. In other words, data citation is part of the open science movement because rewarding credit to those people who share their data is essential for the movement to maintain its momentum. The results of this study indicate, however, that 90% of references to data do not conform to traditional citation practices. Accordingly, reliable measurement of the impact of open research data and careful consideration of the ways in which, and extent to which, open research data are shared, reused, and hopefully, cited are essential going forward. The findings presented here demonstrate that an increase has occurred in data sharing and that dramatic differences exist among disciplines, facts that publishers of journals and decision makers at higher education institutions and funding agencies need to keep in mind when developing guidelines, recommendations, policies and standards for data citation.

# REFERENCES

Aboelela, S. W., Larson, E., Bakken, S., Carrasquillo, O., Formicola, A., Glied, S. A., . . . Gibbie, K. M. (2007). Defining interdisciplinary research: Conclusions from a critical review of the literature. *Health Services Research, 42*, 329-346. https://doi.org/10.1111/j.1475-6773.2006.00621.x

Ajiferuke, I., Lu, K., & Wolfram, D. (2010). A comparison of citer and citation-based measure outcomes for multiple disciplines. *Journal of the American Society for Information Science and Technology, 61*(10), 2086–2096. https://doi.org/10.1002/asi.21383

American Psychological Association. (1983). *Publication manual of the American Psychological Association* (3rd ed.). Washington, DC: American Psychological Association.

Anagnostou, P., Capocasa, M., Milia, N., Sanna, E., Battaggia, C., Luzi, D., & Destro Bisol, G. (2015). When data sharing gets close to 100%: What human paleogenetics can teach the open science movement. *PLoS ONE, 10*(2), e0121409. https://doi.org/10.1371/journal.pone.0121409

Andreoli-Versbach, & Mueller-Langer. (2014). Open access to data: An ideal professed but not practised. *Research Policy, 43*(9), 1621-1633. https://doi.org/10.1016/j.respol.2014.04.008

Archambault, E., & Larivière, V. (2009). History of the journal impact factor: Contingencies and consequences. *Scientometrics, 79*(3), 635-649. https://doi.org/10.1007/s11192-007-2036-x

Aziz, M., & North, S. (2007). *Retrieving software component using clone detection and programing slicing.* Sheffield, UK: The University of Sheffield.

Ball, A. (2009). *Scientific data application profile scoping study report.* Retrieved June 26, 2017, from http://www.ukoln.ac.uk/projects/sdapss/papers/ball2009sda-v11.pdf

Bassett, P. G. (1997). *Framing software reuse: Lessons from the real world.* Upper Saddle River: Yourdon Press.

Bechhofer, S., Buchan, I., Roure, D. D., Missier, P., Ainsworth, J., Bhagat, J., . . . Goble, C. (2013). Why linked data is not enought for scientists. *Future Generation Computer Systems, 29*(2), 599-611. https://doi.org/10.1016/j.future.2011.08.004

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS One, 9*(3), e92590. https://doi.org/10.1371/journal.pone.0092590

Bhattacharya, S., & Basu, P. K. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics, 43*(3), 359-372. https://doi.org/10.1007/BF02457404

Bishop, L. (2005). Protecting respondents and enabling data sharing: Reply to Parry and Mauthner. *Sociology*, 333-336. https://doi.org/10.1177/0038038505050542

Bishop, L. (2009). Ethical sharing and reuse of qualitative data. *Australian Journal of Social Issues, 44*(3), 255-272. https://doi.org/10.1002/j.1839-4655.2009.tb00145.x

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216-1227. https://doi.org/10.1002/asi.20077

Blumenthal, D., Campbell, E. G., Gokhale, M., Yucel, R., Clarridge, B., Hilgartner, S., & Holtzman, N. A. (2006). Data withholdings in genetics and the other life sciences: Prevalences and predictors. *Academic Medicine, 81*(2), 137-145.

Bohlin, L., Edler, D., Lancichinetti, A., & Rosvall, M. (2014). Community detection and visualization of networks with the map equation framework. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: methods and practice* (pp. 3-34). New York:Springer.

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., & Balakireva, L. (2009). Clickstream data yields high-resolution maps of science. *PLoS ONE*, *4*, e4803. https://doi.org/10.1371/journal.pone.0004803

Borgman, C. L. (2007). *Scholarship in the digital age: Information, infrastructure, and the internet.* Cambridge, MA, USA: MIT Press.

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. (2012). Why are the attribution and citation of scientific data important. In P. F. Uhlir (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 1-10). Washington, D.C.: National Academies Press.

Borgman, C. L. (2016). Data citation as a bibliometric oxymoron. In C. R. Sugimoto (Ed.), *Theories of informetrics and scholarly communication* (pp. 93-115). Berlin/Boston: De Gruyter.

Börner, K., Chen, C., & Boyak, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179-255. https://doi.org/10.1002/aris.1440370106

Borouch, R. F. (1985). *Definitions, products, distinctions in data sharing.* (S. E. Fienberg, M. E. Martin, & M. L. Straf, Eds.) Washington, DC, USA: National Academy Press.

Boulton, G., Campbell, P., Collins, B., Hall, D. W., Laurie, G., O'Neill, O., . . . Walport, M. (2012, June). *Science as an open enterprise: The royal society science policy centre report.* Retrieved April 13, 2018, from https://royalsociety.org/topics-policy/projects/science-public-enterprise/Report/

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404. https://doi.org/10.1002/asi.21419

Boyak, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351-374. https://doi.org/10.1007/s11192-005-0255-6

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology, 57*(8), 1060-1072. https://doi.org/10.1002/asi.20373

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science, 37*(1), 34-36. https://doi.org/10.1002/(SICI)1097-4571(198601)37:1<34::AID-ASI5>3.0.CO;2-0

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., . . . Wright, D. (2012). Making data a first class scientific output: Data citation and publication by NERC's environmental data centres. *International Journal of Digital Curation, 7*(1), 107-113. https://doi.org/10.2218/ijdc.v7i1.218

Callon, M., Courtial, J., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research - the case of polymer chemistry. *Scientometrics, 22*(1), 155-205. https://doi.org/10.1007/BF02019280

Campbell, E. G., & Bendavid, E. (2003). Data-sharing and data-withholding in genetics and the life science: Results of a national survey of technology transfer officers. *Journal of Health Care Law Policy, 6*, 241-255.

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology, 66*(9), 1747-1762. https://doi.org/10.1002/asi.23358

Canham, S., & Ohmann, C. (2016). A metadata schemea for data objects in clinical research. *Trials, 17*(1), 557. https://doi.org/10.1186/s13063-016-1686-5

Chao, T. C. (2011). Disciplinary reach: Investigating the impact of dataset reuse in the earth sciences. *Proceedings of the American Society for Information Science and Technology*, (pp. 1-8). https://doi.org/10.1002/meet.2011.14504801125

Chao, T. C. (2015). Mapping methods metadata for research data. *International Journal of Digital Curation, 10*(1), 82-94. https://doi.org/10.2218/ijdc.v10i1.347

Choudhury, S., Fishman, J. R., McGowan, M. L., & Juengst, E. T. (2014). Big data, open science and the brain: lessons learned from genomics. *Frontiers in Human Neuroscience, 8*(239), 1-10. https://doi.org/10.3389/fnhum.2014.00239

Clarivate Analytics. (2012). *Research Areas*. Retrieved January 12, 2017, from https://images.webofknowledge.com/WOKRS57B4/help/WOS/hp_research_areas_easca.html.

Clarivate Analytics. (2016). *Data Citation Index - Resesearch area*. Retrieved January 12, 2017, from http://images.webofknowledge.com/WOKRS523_2R2/help/DRCI/hp_research_areas_easca.html

Clarivate Analytics. (2018). *ESI Master Journal Lilst 2018.* Retrieved June 13, 2018, from http://ipscience-help.thomsonreuters.com/incitesLiveESI/ESIGroup/overviewESI/esiJournalsList.html.

Clarivate Analytics. (2018). *Web of Science core collection help*. Retrieved June 8, 2018, from https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html.

Clarivate Analytics. (2018). *Web of Science platform: Data Citation Index.* Retrieved August 23, 2018, from https://clarivate.libguides.com/webofscienceplatform/dci.

Clark, K. E. (1954). The APA study of psychologists. *American Psychologist, 9*(3), 117-120.

Cliggett, L. (2013). Qualitative data archiving in the digital age: Strategies for data preservation. *The Qualitative Report, 18*(24), 1.

Clubb, J. A., Austin, E. W., Geda, C. L., & Traugott, M. W. (1985). Sharing research data in the social sciences. In S. E. Fienberg, M. E. Martin, & M. L. Straf (Eds.), *Sharing research data* (pp. 39-80). Washington, DC: National Academic Press.

Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E., & Herrerea, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382-1402. https://doi.org/10.1002/asi.21525

CODATA-ICSTI Task Group on Data Citation Practices. (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. (Y. M. Socha, Ed.) *Data Science Journal*, CIDCR1-CIDCR75. https://doi.org/10.2481/dsj.OSOM13-043

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1995). Share and share alike isn't always the rule in science. *Science, 268*(5218), 1715-1718. https://doi.org/10.1126/science.7792594

Costa, M., Qin, J., & Bratt, S. (2016). Emergence of collaboration networks around large scale data repositories: A study of the genomics community using GenBank. *Scientometrics, 108*(1), 21-40. https://doi.org/10.1007/s11192-016-1954-x

Costas, R., Meijer, I., Zahedi, Z., & Wouters, P. F. (2013). *The value of research data metrics from a cultural and technical point of view.* Retrieved February 15, 2018, from http://repository.jisc.ac.uk/6205/1/Value_of_Research_Data.pdf.

Couture, J. L., Blake, R. E., McDonald, G., & Ward, C. L. (2018). A funder-imposed data publication requirement seldom inspired data sharing. *PLoS ONE, 13*(7), e0199789. https://doi.org/10.1371/journal.pone.0199789

Craig, I. D., Plume, A. M., McVeigh, M. E., Pingle, J., & Amin, M. (2007). Do open access article have greater citation impact? : A critical review of the literature. *Journal of Informetrics, 1*(3), 239-248. https://doi.org/10.1016/j.joi.2007.04.001

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities.* Chicago: The University of Chicago Press.

Creswell, J., & Clark, P. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage publications.

Cronin, B. (1984). *The citation process. The role and significance of citations in scientific communication.* London: Taylor Graham.

Cronin, B. (1995). *The scholar's courtesy: The role of acknolwedgement in the primary communication process.* London: Taylor Graham.

Cronin, B. (2001). Acknowledgement trends in the research literature of information science. *Journal of Documentation, 57*(3), 427-433. https://doi.org/10.1108/EUM0000000007089

Cronin, B. (2014). Scholars and scripts, spoors and scores. In B. Cronin, & C. R. Sugimoto (Eds.), *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact* (pp. 3-21). Cambridge, MA: MIT Press.

Curty, R. (2015). *Beyond "data thrifting": An investigation of factors influencing research data reuse in the social sciences.* Retrieved from ProQuest Dissertations & Theses Global. (3713677)

Curty, R. G. (2016). Factors influencing research data reuse in the social sciences: An exploratory study. *International Journal of Digital Curation,* 96-117. https://doi.org/10.2218/ijdc.v11i1.401

Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. *PLoS ONE, 12*(12), e0189288. https://doi.org/10.1371/journal.pone.0189288

Dallmeier-Tiessen, S., Darby, R., Gitmans, K., Lambert, S., Mattews, B., Mele, S., . . . Wilson, M. (2014). Enabling sharing and reuse of scientific data. *New Review of Information.* https://doi.org/10.1080/13614576.2014.883936

Daniels, G. M. (2014). *Data reuse in museum contexts: Experiences of archaeologists and botanists.* Retrieved from ProQuest Dissertations & Theses Global. (3636549)

Darch, P. T., Borgman, C. L., Traweek, S., Cummings, R. L., Wallis, J. C., & Sands, A. E. (2015). What lies beneath?: Knowledge infrastructure in the subseafloor biosphere and beyond. *Internatioinal Journal on Digital Libraries, 16*(1), 61-77. https://doi.org/10.1007/s00799-015-0137-3

Data Citation Synthesis Working Group. (2014). *Joint declaration of data citation principles-Final.* (M. Martone, Ed.) San Diego, CA, USA: FORCE11.

DataCite. (2018). *DataCite Event Data*. Retrieved November 23, 2018, from https://www.datacite.org/eventdata.html

DataCite Metadata Working Group. (2016). *DataCite metadata schema documentation for the publication and citation of research data version 4.0.* http://doi.org/10.5438/0012.

Data Observation Network for Earth. (2018). *DataOne*. Retrieved November 23, 2018, from https://search.dataone.org/data

Davis, P. M. (2010). Does open access lead to increased readership and citations? A randomized controlled trial of articles published in APS journals. *The Physiologist, 53*(6), 200-201. https://doi.org/10.1096/fj.11-183988

Diamantopoulos, N., Sgouropoulou, C., Kastrantas, K., & Manouselis, N. (2011). Developing a metadata application profile for sharing agricultural scientific and scholarly research resources. *Research Conference on Metadata and Semantic Research* (pp. 453-466). Springer.

Digital Curation Center. (2018). *Disciplinary metadata.* Retrieved July 24, 2018, from http://www.dcc.ac.uk/resources/metadata-standards.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management, 37*(6), 817-842. https://doi.org/10.1016/S0306-4573(00)00051-0

Dodd, S. A. (1979). Bibliographic references for numeric social science data files: Suggested guidelines. *Journal of the American Society for Information Science, 30*(2), 77-82. https://doi.org/10.1002/asi.4630300203

Drachen, T. M., Elleggard, O., Larsen, A. V., & Fabricius, S. V. (2016). Sharing data increases citations. *The Journal of European Research Libraries*, *26*(2), 67-82. http://doi.org/10.18352/lq.10149

Dranchen, T. M., Ellegaard, O., Larsen, A. V., & Dorch, S. (2016). Sharing data increases citations. *LIBER Quarterly*, *26*(2), 67-82. http://doi.org/10.18352/lq.10149

Economic and Social Research Council. (2015). *Research Data Policy*. Retrieved May 9, 2017, from http://www.esrc.ac.uk/funding/guidance-for-grant-holders/research-data-policy/

Egghe, L. (2005). Expansion of the field of informetrics: Origins and consequences. *Information Processing and Management*, *41*(6), 1311-1316. https://doi.org/10.1016/j.ipm.2005.03.011

Egghe, L., & Rousseau, R. (1996). Averaging and globalising quotients of informetric and scientometric data. *Journal of Information Science*, *22*(3), 165-170. https://doi.org/10.1177/016555159602200302

European Commission. (2011). *Digital agenda: Turning government data into gold.* Retrieved April 13, 2018, from http://europa.eu/rapid/press-release_IP-11-1524_en.htm

European Commission. (2016). *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020.* Retrieved May 10, 2017, from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Executive Office of the President. (2013). *Open data policy-Managing information as an asset.* Retrieved from https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

Eysenbach, G. (2006). Citation advantage of open access articles. *PLOS Biology, 4*(5), e157. https://doi.org/10.1371/journal.pbio.0040157

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Journal of Computer Supported Cooperative Work, 19*(3-4), 355-375. https://doi.org/10.1007/s10606-010-9117-8

Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data reuse and sensemaking among novice social scientists. *Proceedings of the Association for Information Science and Technology, 49,* pp. 1-10. https://doi.org/10.1002/meet.14504901068

Fear, K. M. (2013). *Measuring and anticipating the impact of data reuse*. Ann Arbor, MI, USA: University of Michigan-Ann Arbor.

Feinberg, S. E., Martin, M. E., & Straf, M. L. (1985). *Sharing research data.* National Academy Press.

Fenner, M., Lowenberg, D., Jones, M., Needham, P., Vieglais, D., Abrams, S., … Chodacki J. (2018). Code of practice for research data usage metrics release 1. *PeerJ Preprints,* 6:e26505v1. https://doi.org/10.7287/peerj.preprints.26505v1

García, F., Bertoa, M. F., Calero, C., Vallecillo, A., Ruíz, F., Piattini, M., & Genero, M. (2006). Towards a consistent terminology for software measurement. *Information and Software Technology, 48*(8), 631-644. https://doi.org/10.1016/j.infsof.2005.07.001

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(3159), 108-111. https://doi:10.1126/science.122.3159.108

Garfield, E. (2006). The history and meaning of the Journal Impact Factor. *Journal of the American Medical Association, 295*(1), 90-93. https://doi:10.1001/jama.295.1.90

Garfield, E., Malin, M. V., & Small, H. (1978). Citation datda as science indicators. In J. Elkana, J. Lederberg, R. K. Merton, A. Thackray, & H. Zuckerman (Eds.), *Toward a metric of science: The advent of science indicators.* New York: Wiley.

Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation, 14*(3), 195-201. https://doi.org/10.1002/asi.5090140304

Gaston, J. (1978). *The reward system in British and American science.* New York: Wiley.

Gilbert, G. N. (2015). Referencing as persuasion. *Social Studies of Science, 7*, 113-122. https://doi.org/10.1177/030631277700700112

Glänzel, W., & Thijs, B. (2004). Does co-authorship inflate the share of self-citations? *Scientometrics, 61*(3), 395-404. https://doi.org/10.1023/B:SCIE.0000045117.13348.b1

Gómez, I., Bordons, M., Fernández, M. T., & Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics, 35*(2), 223-235. https://doi.org/10.1007/BF02018480

Gordon, I. E., Potterbusch, M. R., Bouquin, D., Erdmann, C. C., Wilzewski, J. S., & Rothman, L. S. (2016). "Are your spectroscopic data being used?". *Journal of Molecular Spectroscopy, 327*, 232-238. https://doi.org/10.1016/j.jms.2016.03.011

Google. (2018). *Google Dataset Search*. Retrieved November 23, 2018, from https://toolbox.google.com/datasetsearch.

Gravetter, F. J., & Forzano, L.-A. B. (2012). *Research methods for the behavioral sciences* (4 ed.). Belmont, CA, Australia: Wadsworth Cengage Learning.

Greenfield, P., Droettboom, M., & Bray, E. (2015). ASDF: A new data format for astronomy. *Astronomy and Computing, 12*, 240. https://doi.org/10.1016/j.ascom.2015.06.004

Grosbol, P. (1988). The FITS Data Format. *Bulletin d'Information du Centre de Donnees Stellaires, 35*, 7.

Gross, P. L., & Gross, E. M. (1927). College libraries and chemical education. *Science, 66*(1713), 385-389. https://doi:10.1126/science.66.1713.385

Harnad, S., & Brody, T. (2004). Comparing the impact of Open Access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine, 10*(6). Retrieved November 16, 2016, from http://www.dlib.org/dlib/june04/harnad/06harnad.html

Harter, S. P., & Nisonger, T. E. (1997). ISI's impact factor as misnomer: A proposed new measure to assess journal impact. *Journal of the American Society for Information Science, 48*(12), 1146-1148. https://doi.org/10.1002/(SICI)1097-4571(199712)48:12<1146::AID-ASI9>3.0.CO;2-U

He, L., & Nahar, V. (2016). Reuse of scientific data in academic publications: An investigation of Dryad digital repository. *Aslib Journal of Information Management*, *68*(4), 478-494. https://doi.org/10.1108/AJIM-01-2016-0008

Helbig, K., Hausstein, B., & Toepfer, R. (2015). Supporting data citation: Experiences and best practices of a DOI allocation agency for social sciences. *Journal of Librarianship and Scholarly Communication, 3*(2), eP1220. http://doi.org/10.7710/2162-3309.1220

Henry, E., & Faller, B. (1995). Large-scale industrial reuse to reduce cost and cycle time. *Software, IEEE*, *12*(5), 47-53. https://doi.org/10.1109/52.406756

Hinds, P. S., Vogel, R. J., & Clarke-Steffen, L. (1997). The possibilities and pitfalls of doing a secondary analysis of a qualitative data sets. *Qualitative Health Research, 7*(3), 408-424. https://doi.org/10.1177/104973239700700306

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science, 46*(6), 400-425. https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Hong, N. C. (2014). Minimal information for reusable scientific software. *Proceedings of the 2nd Workshop on Working towards Sustainable Scientific Software: Practice and Experience*. https://doi.org/10.1101/429142

Hourclé, J., Chang, W., Linares, F., & Palanisamy, G. (2012). *Linking articles to data.* Retrieved July 24, 2018, from https://vso1.nascom.nasa.gov/rdap/RDAP2012_landingpages_handout.pdf

Howison, J., & Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology lilterature. *Journal of the Association for Information Science and Technology, 67*(9), 2137-2155. https://doi.org/10.1002/asi.23538

Howison, J., & Herbsleb, J. D. (2011). Scientific software production: Incentives and collaboration. *Proceedings of the ACM 2011 conference on Computer Supported Cooperative Work*, (pp. 513-522). https://doi.org/10.1145/1958824.1958904

Infrastructure Service for Open Access. (2018). *Directory of Open Access Journals*. Retrieved February 26, 2018, from https://doaj.org/

Israel, M., & Hay, I. (2006). *Research ethics for social scientists.* London: Sage.

Jarneving, B. (2005). A comparison of two bibliometric methods for mapping of the research front. *Scientometrics, 65*(2), 245-263. https://doi.org/10.1007/s11192-005-0270-7

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics, 1*(4), 287-307. https://doi.org/10.1016/j.joi.2007.07.004

Jeng, W. (2017). *Qualitative data sharing practices in social sciences.* Retrieved from ProQuest Dissertations & Theses Global. (10645840)

Jirotka, M., Procter, R., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., . . . Voss, A. (2005). *Collaboration and trust in healthcare innovation: The eDiaMoND case study.* Dordrecht, The Netherlands: Kluwer.

Joo, S., Kim, S., & Kim, Y. (2017). An exploratory study of health scientists' data reuse behaviors: Examining attitudinal, social, and resource factors. *Aslib Journal of Information Management, 69*(4), 389-407. https://doi.org/10.1108/AJIM-12-2016-0201

Joo, Y. K., & Kim, Y. (2017). Engineering researchers' data reuse behaviors: A structural equation modeling approach. *The Electronic Library, 35*(6), 1141-1161. https://doi.org/10.1108/EL-08-2016-0163

Katz, D. S., & Smith, A. M. (2015). Transitive credit and JSON-LD. *Journal of Open Research Software, 3,* e7. http://doi.org/10.5334/jors.by

Katz, D. S., Niemeyer, K. E., Smith, A. M., Anderson, W. L., Boettiger, C., Hinsen, K., . . . Rios, F. (2016). Software vs. data in the context of citation. *PeerJ Preprints*. https://doi.org/10.7287/peerj.preprints.2630v1

Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics: Re-sharing scientific practice. *Nature Reviews Genetic, 10*(5), 331-335. https://doi.org/10.1038/nrg2573

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10-25. https://doi.org/10.1002/asi.5090140103

Kim, Y. (2013). *Institutional and individual influence on scientists' data sharing behaviors.* Retrieved from ProQuest Dissertations & Theses Global.(3568789).

Kim, Y., & Adler, M. (2015). Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. *International*

*Journal of Information Management*, *35*(4), 418-418. https://doi.org/10.1016/j.ijinfomgt.2015.04.007

Kim, Y., & Stanton, J. M. (2015). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology, 67*(4), 776-799. https://doi.org/10.1002/asi.23424

Kim, Y., & Yoon, A. (2017). Sceintists' data reuse behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology, 68*(12), 2709-2719. https://doi.org/10.1002/asi.23892

Kim, Y., & Zhang, P. (2015). Understanding data sharing behaviors of STEM researchers: The role of attitudes, norms, and data repositories. *Library and Information Science Research, 37*(3), 189-200. https://doi.org/10.1016/j.lisr.2015.04.006

King, G. (1995). Replication, replication. *PS: Political Science and Politics, 28*(3), 444-452.

https://doi.org/10.2307/420301

Klein, J. T. (1990). *Interdisciplinary: History, theory, and practice.* Detroit, Michigan: Wayne State University.

Kling, R., & Spector, L. (2003). Rewards for scholarly communication. In D. L. Andersen (Ed.), *Digital scholarship in the tenure, promotion, and review process* (pp. 78-103). Armonk, NY: ME Sharpe, Inc.

Krauss, J. (2007). Journal self-citation rates in ecological sciences. *Scientometrics*, *73*(1), 79-89. https://doi.org/10.1007/s11192-007-1727-7

Krueger, C. (1992). Software reuse. *ACM Computing Surveys, 24*(2), 131-183.

Kruskal, J. K. (1964). Multidimensiional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1-27. https://doi.org/10.1007/BF02289565

Lagoze, C. (2000). *Accommodating simplicity and complexity in metadata: Lessons from the Dublin Core experience.* Retrieved June 26, 2017, from https://ecommons.cornell.edu/handle/1813/5792

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* Cambridge, MA, USA: Harvard University Press.

Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics, 23*(3), 417-461. https://doi.org/10.1007/BF02029807

Lawani, S. M., & Bayer, A. E. (1983). Validity of citation criteria for assessing the influence of scientific publications: New evidence with peer assessment. *Journal of the American Society for Information Science, 34*(1), 59-66. https://doi.org/10.1002/asi.4630340109

Lawrence, B., Jones, C., Mattews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation, 6*(2), 4-37. https://doi.org/10.2218/ijdc.v6i2.205

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: The importance of species similarity. *Ecology, 93*(3), 477-489. https://doi.org/10.1890/10-2402.1

Lercher, A. (2013). Correlation over time for citatons to mathematics articles. *Journal of the Association for Information Science and Technology, 64*(3), 455-463. https://doi.org/10.1002/tea.3660180403

Leydesdorff, L. (1997). Why words and co-words cannot map the development of the science. *Journal of the American Society for Information Science*, *48*(5), 418-427. https://doi.org/10.1002/(SICI)1097-4571(199705)48:5<418::AID-ASI4>3.0.CO;2-Y

Leydesdorff, L. (2008). Caveats for the use of citation indicators in research and journal evaluations. *Journal of the Association for Information Science and Technology*, *59*(2), 279–297. https://doi.org/10.1002/asi.20743

Leydesdorff, L. (2018). Diversity and interdisciplinarity: How can one distinguish and recombine disparity, variety, and balance? *Scientometrics*, *116*(3), 2113-2121. http://doi.org/10.1007/s11192-018-2810-y.

Leydesdorff, L., de Moya-Anegón, F., & Guerrero-Bote, V. (2010). Journal maps on the basis of Scopus data: A comparison with the Journal Citation Reports of the ISI. *Journal of the American Society for Information Science and Technology*, *61*(2), 352-369. https://doi.org/10.1002/asi.21250

Li, K., Greenberg, J., & Lin, X. (2016). Software citation, reuse and metadata considerations: An exploratory study examining LAMMPS. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1-10. https://doi.org/10.1002/pra2.2016.14505301072

Li, K., Yan, E., & Feng, Y. (2017). How is R cited in research outputs? Structure, impacts, and citation standard. *Journal of Informetrics*, *11*(4), 989-1002. http://doi.org/10.1016/j.joi.2017.08.003.

Lin, J. (2009). Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, *10*(1), 46. https://doi.org/10.1186/1471-2105-10-46

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology, 63*(10), 1973-1986. https://doi.org/10.1002/asi.22628

Lu, K., Ajiferuke, I., & Wolfram, D. (2014). Extending citer analysis to journal impact evaluation. *Scientometrics, 100*(1), 245-260. https://doi.org/10.1007/s11192-014-1274-y

Marcus, A., & Menzies, T. (2010). Software is data too. *Proceedings of the FSE/SDP workshop on future of software engineering research* (pp. 229-232). New York: ACM. https://doi.org/10.1145/1882362.1882410

Marshakova, I. V. (1973). System of document connectios based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy, 6*, 3-8.

Mason, J. (2007). 'Re-Using' qualitative data: on the merits of an investigative epistemology. *Sociological Resesarch Online, 12*(3), 3. https://doi.org/10.5153/sro.1507

Mauthner, N. (2012). Are research data a' Common' resource? *feminists@law*. Retrieved February 26, 2018, from https://journals.kent.ac.uk/kent/index.php/feministsatlaw/article/view/60

Mauthner, N. S., & Parry, O. (2009). Qualitative data preservation ahd sharing in the social sciences: On whose philosophhical terms? *Australian Journal of Social Issues, 44*(3), 289-305. https://doi.org/10.1002/j.1839-4655.2009.tb00147.x

McIlroy, M. D. (1968). Mass produced software components. In P. Naur, & B. Randell (Ed.), *In Software Engineering*, (pp. 138-150).

Merton, R. K. (1968). The Mattew effect in science. *Science, 159*(3810), 56-63. https://doi.org/10.1126/science.159.3810.56

Merton, R. K. (1988). The Mattew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis, 79*(4), 606-623.

Mockus, A. (2007). Large-scale code reuse in open-source software. *International Workshop on Emerging Trends in FLOSS Research and Development*, (p. 0:7). https://doi.org/10.1109/FLOSS.2007.10

Moed, H. F. (2005). *Citation analysis in research evaluation.* Dordrecht: Springer.

Moed, H. F., Glänzel, W., & Schmoch, U. (2004). *Handbook of Quantitative Science and Technology Research.* New York: Kluwer Academic Publishers.

Moed, H. F., Van Leeuwen, T. N., & Reedijk, J. (1999). Towards appropriate indicators of journal impact. *Scientometrics, 46*(3), 575-589. https://doi.org/10.1007/BF02459613

Mongeon, P., Robinson-García, N., Jeng, W., & Costas, R. (2017). Incorporating data sharing to the reward system of science: Linking DataCite records to authors in the Web of Science. *Aslib Journal of Information Management*. https://doi.org/10.1108/AJIM-01-2017-0024

Mooney, H. (2011). Citing data sources in the social sciences: Do authors do it? *Learned Publishing, 24*(2), 99-108. https://doi.org/10.1087/20110204

Mooney, H., & Newton, M. P. (2012). The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianschip and Scholarly Communication, 1*(1), eP1035. https://doi.org/10.7710/2162-3309.1035

Morisio, M., Ezran, M., & Tully, C. (2002). Success and failure factors in software reuse. *IEEE Transaction on Software Engineering, 28*(4), 340-357. https://doi.org/10.1109/TSE.2002.995420

Moya-Anegon, F., Vargas-Quesada, B., Chinchilla-Rodriguez, Z., Corera-Alvarez, Herrero-Solana, V., & Munoz-Fernandez, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics, 61*(1), 129-145. https://doi.org/10.1023/B:SCIE.0000037368.31217.34

Nangia, U., & Katz, D. S. (2017). Track 1 Paper: Surveying the U.S. National Postdoctoral Association regarding software use and training in research. *Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE 5.2).* https://doi.org/10.5281/zenodo.814220

National Institutes of Health. (2003). *NIH data sharing policy and implementation guidance*. Retrieved January 15, 2017, from https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

National Science Foundation. (2010). *Instructions and codes for completing project data form (Form 1295).* Retrieved January 12, 2017, from https://nsf.gov/pubs/2010/nsf10034/nsf10034.docx

National Science Foundation. (2011). *Digital research data sharing and management.* Retrieved January 15, 2017, from www.nsf.gov/nsb/publications/2011/nsb1124.pdf

National Science Foundation. (2013). *Grant proposal guidelines: NSF 13-1 January 2013, GPG summary of changes*. Retrieved November 6, 2017, from https://nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_sigchanges.jsp

Nicolescu, B. (1998). *The transdisciplinary evolution of learning.* Retrieved April 9, 2018, from https://pdfs.semanticscholar.org/8ff2/bde0bd4dd47bf2c4cd5396be6e90221df786.pdf

Nijssen, D., Rousseau, R., & van Hecke, P. (1998). The Lorenz curve: a graphical representation of evenness. *Coenoses, 13*(1), 33-38.

Niu, J. (2009). *Perceived documentation quality of social science data.* Retrieved from ProQuest Dissertations & Theses Global. (3382314).

Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the Association for Information Science and Technology, 59*(12), 1963-1972. https://doi.org/10.1002/asi.20898

Pan, X., Yan, E., & Hua, W. (2016). Disciplinary differences of software use and impact in scientific lilterature. *Scientometrics, 109*(3), 1593-1610. https://doi.org10.1007/s11192-016-2138-4

Park, H., & Wolfram, D. (2017). An Examination of Research Data Sharing and Re-Use: Implications for Data Citation Practice. *Scientometrics, 111*(1), 443-461. https://doi.org/10.1007/s11192-017-2240-2

Park, H., You, S., & Wolfram, D. (2017). Is informal data citation for data sharing and re-use more common than formal data citation? *Proceedings of the Association for Information Science and Technology 54.*(1), 768-769. https://doi.org/10.1002/pra2.2017.14505401150

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24049

Parry, O., & Mauthner, N. S. (2004). Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. *Sociology, 38*(1), 139-152. https://doi.org/10.1177/0038038504039366

Parsons, M. A., Godøy, Ø., LeDrew, E., De Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science, 37*(6), 555-569. https://doi.org/10.1177/0165551511412705

Parsons, M. A., Ruth, D., & Minster, J.-B. (2010). Data Citation and Peer Review. *Eos, Transactions American Geophysical Union, 91*(34), 297-298. https://doi.org/10.1029/2010EO340001

Peiling, W., You, S., Manasa, R., & Wolfram, D. (2016). Open peer review in scientific publishg: A web mining study of PeerJ authors and reviewers. *Journal of Data and Information Science, 1*(4), 60-80. https://doi.org/10.20309/jdis.201625

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. *PLoS ONE, 9*(8), e104798. https://doi.org/10.1371/journal.pone.0104798

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2015). Research Data Explored: Citations versus Altmetrics. *15th International Conference on Scientometrics and Informetrics*, (pp. 172-183).

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: An extended analysis of citations and altmetrics. *Scientometrics, 107*(2), 723-744. https://doi.org/10.1007/s11192-016-1887-4

Peters, M. A., & Roberts, P. (2012). *The virtues of openness: Education, science, and scholarship in the digital age.* Boulder, CO: Paradigm Publishers.

Pierce, S. J. (1999). Boundary crossing in research literature as a means of interdisciplinary information transfer. *Journal of the American Society for Information Science, 50*(3), 271-279. https://doi.org/10.1002/(SICI)1097-4571(1999)50:3<271::AID-ASI10>3.0.CO;2-M

Piwowar, H. A. (2010). *Foundational studies for measuring the impact, prevalence, and patterns of publicly sharing biomedical research data*. Retrieved from ProQuest Dissertations & Theses Global. (3417405)

Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE, 6*(7), e18657. https://doi.org/10.1371/journal.pone.0018657.g003

Piwowar, H. A., & Chapman, W. W. (2010). Public sharing of research datasets: A pilot study of associations. *Journal of Informetrics, 42*(2), 148-156. https://doi.org/10.1016/j.joi.2009.11.010

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ, 1*, e175. https://doi:10.7717/peerj.175

Piwowar, H. A., Day, R., & Fridsma, D. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE*, *2*(3), e308. https://doi.org/10.1371/journal.pone.0000308

Poist, T. (2015). Best publishing practices to improve user confidence in scientific software. *Ideas in Ecology and Evolution, 8*, 50-55. https://doi.org/10.1371/journal.pbio.1001745

Priem, J., & Costello, K. (2010). How and why scholars cite on Twitter. *Proceedings of the American Society for Information Science and Technology*, *47*, 1-4. https://doi.org/10.1002/meet.14504701201

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation, 25*(4), 348-349.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinary: Case study in bionanoscience. *Scientometrics, 82*(2), 263-287. https://doi.org/10.1007/s11192-009-0041-y

Rasmussen, K. B. (2011). Barking up the right tree. *IASSIST Quarterly*.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2 ed.). Thousand Oaks: Sage Publications.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science, 331*(6018), 703-705. https://doi:10.1126/science.1197962

Reidpath, D. D., & Allotey, P. A. (2001). Data sharing in medical research: An empirical investigation. *Bioethics*(15), 125-134. https://doi.org/10.1111/1467-8519.00220

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology, 67*(12), 2964-2975. https://doi.org/10.1002/asi.23529

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics, 11*(3), 841-854. https://doi.org/10.1016/j.joi.2017.07.003

Rolland, B., & Lee, C. (2013). Beyond trust and reliability: Reusing data in collaborative cancer epidemiology researCh. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, (pp. 435-444). https://doi.org/10.1145/2441776.2441826

Rowe, B. R., Wood, D. W., Link, A. N., & Simoni, D. A. (2010). *NIST's text retrieval conference (TREC) program.* Retrieved August 25, 2018, from https://trec.nist.gov/pubs/2010.economic.impact.pdf

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE, 4*(9), e7078. https://doi.org/10.1371/journal.pone.0007078

Schubert, A., & Braun, T. (1993). Reference standards for citation based assessments. *Scientometrics*, *26*(1), 21-35. https://doi.org/10.1007/BF02016790

Schwartz, C. A. (1997). The rise and fall of uncitedness. *College & Research Libraries, 58*(1), 19-29. https://doi.org/10.5860/crl.58.1.19

Seglen, P. O. (1994). Casual relationship between article citedness and journal impact. *Journal of the American Society for Information Science, 45*(1), 1-11. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<1::AID-ASI1>3.0.CO;2-Y

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6-20. https://doi.org/10.1002/asi.23917

Slavnic, Z. (2011). *Preservation and sharing of qualitative data-academic debate and policy developments.* Linköping: REMESO.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology, 43*(1), 1-43. https://doi.org/10.1002/aris.2009.1440430113

Small, H. (1973). Co-citation in the Scientific Literature: A New measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science, 24*(4), 265-269. https://doi.org/10.1002/asi.4630240406

Smiraglia, R. P. (2002). Further progress toward theory in knowledge organization. *Canadian Journal of Information and Library Science, 26*(2-3), 30-49.

Smiraglia, R. P. (2012). *Domain analysis for knowledge organization: Tools for ontology extraction.* Waldham, MA: Chandos Publishing.

Smiraglia, R. P. (2014). *The elements of knowledge organization.* Cham: Springer International Publishing.

Smith, Arfon M.; Daniel, Katz D.; Niemeyer, Kyle E.; FORCE11 Software Citation Working Group. (2016). Software citation principles. *PeerJ Computer Science,* e.86. https://doi.org/10.7717/peerj-cs.86

Smith, L. C. (1981). Citation analysis. *Library Trends, 30*(1), 83-106.

Stanley, B., & Stanley, M. (1988). Data sharing. The primary researcher's perspective. *Law and Human Behavior, 12*(2), 173-180. http://dx.doi.org/10.1007/BF01073125

Star, J., & Gastl, A. (2011). isCitedBy: A metadata scheme for DataCite. *D-Lib Magazine, 17*(1). https://doi.org/10.1045/january2011-starr

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., . . . Clark, T. (2015). Archiving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science, 1*(1), e1. https://doi.org/10.7717/peerj-cs.1

Sterlling, T. D., & Weinkam, J. J. (1990). Sharing scientific data. *Communications of the ACM, 33*(8), 112-119. https://doi.org/10.1145/79173.79182

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society, Interface, 4*(15), 707-719. https://doi.org/10.1098/rsif.2007.0213

Suber, P. (2002). Open access to the scientific journal literature. *Journal of Biology, 1*(1), 3. https://doi.org/10.1186/1475-4924-1-3

Swauger, S., & Vision, T. J. (2015). What factors influence where researchers deposito their data? A survey of researcher submissions to data repositories. *International Journal of Digital Curation, 10*(1), 68-81. https://doi.org/10.2218/ijdc.v10i1.289

Swoger, B. (2012). Thomson Reuters Data Citation Index. *Library Journal*. Retrieved January 12, 2017, from http://wokinfo.com/media/pdf/dci-libjrnl-review.pdf

Tabachnick, B. G., & Fidell, L. S. (2000). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn and Bacon.

Tague-Sutchliffe, J. (1992). An introduction to informetrics. *Information Processing and Management, 28*(1), 1-3. https://doi.org/10.1016/0306-4573(92)90087-G

Tarnow, E. (2002). Coauthorship in physics. *Science and Engineering Ethics, 8*(2), 175-190. https://doi.org/10.1007/s11948-002-0017-2

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., . . . Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLos ONE, 6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., . . . Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE, 10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Thelwall, M. (2014). *Big data and social web research methods.* University of Wolverhampton.

Thomlinson, R. (1983). Interdisciplinary in the teaching of demomgraphy in North America. *Janasamkhya, 1*, 91-97.

Torres-Salinas, D., Jiménez-Contreras, E., & Robinson-García, N. (2014). How many citations are there in the Data Citation Index? *arXiv preprint*. https://doi.org/arXiv:1409.0753

van Eck, N. J., & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact* (pp. 285-320). Springer.

Van Raan, A. (1998). The influence of international collaboration on the impact of research results : Some simple mathematical considerations concerning the role of self-citations. *Scientometrics, 42*(3), 423-428. https://doi.org/10.1007/BF02458380

Vishwas, C., & Lyubomir, P. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics, 12*(Suppl 15), S2. https://doi.org/10.1186/1471-2105-12-S15-S2

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., . . . Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics, 5*(1), 14-26. https://doi.org/10.1016/j.joi.2010.06.004

Wallis, J. C., & Borgman, C. L. (2011). Who is responsible for data? An exploratory study of data authorship, owndership, and responsibility. *Proceedings of the American Society for Information Science and Technology*, *48*, pp. 1-10. https://doi.org/10.1002/meet.2011.14504801188

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE, 8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Wang, P., Hoyt, J., Pöschl, U., Wolfram, D., Ingwersen, P., Smith, R., & Bates, M. (2016). The last frontier in open science: Will open peer review transform scientific and scholarly publishing? *Proceedings of the Association for Information Science and Technology*, *53*, pp. 1-4. https://doi.org/10.1002/pra2.2016.14505301001

Wang, P., You, S., Rath, M., & Wolfram, D. (2016). Open peer review in scientific publishing: A Web mining study of PeerJ authors and reviewers. *Journal of Data and Information Science*, 1(4), 60-80. https://doi.org/10.20309/jdis.201625

Ward, G., & Baxter, A. (2016). *Distributing Python Modules*. Retrieved August 26, 20187, from https://docs.python.org/3.6/distutils/setupscript.html#additional-meta-data

Weber, N., & Thomer, A. (2014). Paratexts and documentary practices: Text-mining a bioinformatics corpus. In N. Desrochers, & D. Apollon (Eds.), *Examining paratextual theory and its applications in digital culture* (pp. 84-109). Hershey, PA: IGI Global.

White, D. R. (1991). Sharing anthropological data with peers and third world hosts. *Sharing social science data: Advantages and challenges* (pp. 42-61). SAGE Focus Edition.

White, H. D. (1982). Citation analysis of data file use. *Library Trends, 31*(3), 467-477.

White, H. D., & Griffith, B. C. (1980). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, *32*(3), 163-171. https://doi.org/10.1002/asi.4630320302

White, H. D., & McCain, K. W. (1997). Visualization of literatures. *Annual Review of Information Science and Technology, 32*, 99-168.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327-355. https://doi.org/10.1002/(SICI)1097-4571(19980401)49:4<327::AID-ASI4>3.0.CO;2-4

Whitfield, K., & Reid, C. (2004). Assumptions, ambiguities, and possibilities in interdisciplinary population health research. *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique, 95*(6), 434-436. https://www.jstor.org/stable/41994424

Willinsky, J. (2005). The unacknowledged convergence of open source, open access, and open science. *First Monday, 10*(8).

Wilson, C. (1999). Informetrics. *Annual Review of Information Science and Technology, 34*, 107-247. https://doi.org/10.1080/14649055.2003.10765939

Witman, P. D. (2007). *Software reuse "in the Large" - Tracing patterns of reuse.* Claremont, CA, USA: Claremont University.

Wolfram, D. (2003). *Applied informetrics for information retrieval research.* Westport, Conneticut, USA: Libraries Unlimited.

Wouters, P. (1999). *The citation culture.* University of Amsterdam.

Wren, J. D., Kozak, K. Z., Johnso, K. R., Deakyne, S. J., Schilling, L. M., & Dellavalle, R. P. (2007). The write position. *EMBO Reports, 8*(11), 988-991. https://doi:10.1038/sj.embor.7401095

Yoon, A. (2014). "Making a square fit into a circle": Researchers' experiences reusing qualitative data. *Proceedings of the American Society for Information Science and Technology, 51*(1), 1-4. https://doi.org/10.1002/meet.2014.14505101140

Yoon, A. (2015). *Data reuse and users' trust judgments: Toward trusted data curation*. Retrieved from ProQuest Dissertations & Theses Global. (3719920)

Yoon, A. (2017). Data reusers' trust development. *Journal of the Association for Information Science and Technology, 68*(4), 946-956. https://doi.org/10.1002/asi.23730

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account.

*Journal of the Association for Information Science and Technology, 67*(5), 1257-1265. https://doi.org/10.1002/asi.23487

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996-2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology, 59*(13), 2070-2086. https://doi.org/10.1002/asi.20910

Zhao, M., Yan, E., & Li, K. (2018). Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology, 69*(1), 32-46. https://doi.org/10.1002/asi.23919

Zimmerman, A. S. (2008). New knowledge from old data: The rold of standards in the sharing and reuse of ecological data. *Science, Technology, & Human Values, 33*(5), 631-652. https://doi.org/10.1177/0162243907306704

# APPENDICES

## Appendix A. Astronomy/physics: All data types that received at least one data citation (data type, number of total data citation)

- mass spectral data (31,072)
- nmr results (6,157)
- spectral data (3,723)
- software (1,396)
- image file (233)
- fits file (190)
- final output pics (163)
- data (107)
- dataset (63)
- hrcrop (60)
- tex appb (50)
- asc appb (47)
- fesc data (44)
- ascii file (44)
- halo finding (43)
- star data (41)
- anyl files (41)
- fits header file (40)
- photometric calibrations (9)
- fits images (9)
- fits image (9)
- astrometric calibrations (9)
- fileset (8)
- catalog (8)
- images (6)
- ubvri catalog (5)
- still images or photos (5)
- documentation (5)
- asc appa (5)
- fits cube (4)
- ascii spectrum (4)
- text (3)
- model files (3)
- poster (2)
- plot (2)
- photometry (2)
- readme info file (1)
- readme (1)
- rdf (1)
- radio and x ray data (1)
- quantitative data (1)
- processed map data (1)
- presentation (1)
- notebook (1)
- models derived from small angle scattering data (1)
- masks (1)
- mask definition files (1)
- manual (1)
- imaging (1)
- image (1)
- idl sav (1)
- GIS vector data (1)
- fits variables (1)
- fits image gzipped (1)

| | | |
|---|---|---|
| tex figs (38) | idl save files (2) | fits event list (1) |
| ps file (36) | idl pro (2) | file description (1) |
| med (25) | fits files (2) | experimental small angle scattering data from biological macromolecules (1) |
| tex tables (22) | excel (2) | doc (1) |
| paper figs (22) | csv (2) | database (1) |
| gmos pre imaging (20) | textual data individual micro level (1) | code (1) |
| spectra (17) | tex text (1) | catalogs (1) |
| raw data (14) | tapes and transcripts group discussion tape recordings personal documents press clippings minutes of meetings audio cassette tapes (1) | astronomical radio (1) |
| fits header (14) | supplementary materials (1) | astrometry (1) |
| redshifts (10) | spectroscopic data (1) | 2d spectra (1) |
| tex appa (9) | sample data (1) | 1d spectra (1) |

## Appendix B. Biological sciences: All data types that received at least one data citation (data type, number of total data citation)

| | | |
|---|---|---|
| SRA (931,673) | genome binding occupancy profiling by high throughput sequencing (6,118) | methylation profiling by genome tiling array (513) |
| protein sequence data (525,973) | protein structure (4,909) | rnai phenotype data (442) |
| sra (277,920) | fileset (4,406) | methylation profiling by array (383) |
| genomic (163,349) | numeric (3,853) | expression profiling by rt pcr (380) |
| images (113,107) | mixed (3,364) | image stored as reals (379) |
| nucleotide sequencing information (109,135) | datafile (3,008) | snp genotyping by snp array (338) |
| molecular structure (75,899) | sage (2,508) | flow cytometry data (323) |
| processed (72,717) | non coding rna profiling by array (2,380) | dataset unite species hypothesis (300) |
| fgem (72,717) | map (2,235) | expression profiling by sage (230) |
| plant trascription factors and their annotation (65,536) | non coding rna profiling by high throughput sequencing (1,999) | phenotype strain survey (211) |

| | | |
|---|---|---|
| quantitative trait locus map information (55,623) | biosamples (1,893) | mpss (211) |
| adf (55,030) | digital (1,854) | dataset (200) |
| raw (52,496) | genome binding occupancy profiling by genome tiling array (1,804) | case control (186) |
| processed data (48,825) | sequence (1,615) | genome binding occupancy profiling by array (167) |
| sequence data (47,150) | methylation profiling by high throughput sequencing (1,173) | kinomescan (165) |
| raw data (44,591) | datapackage (1,154) | protein profiling by protein array (160) |
| expression profiling by array (40,589) | phylogenetic tree data (978) | third party reanalysis (157) |
| sdrf (40,092) | genome variation profiling by genome tiling array (955) | cel (149) |
| idf (39,365) | annotation (928) | gigadb dataset (124) |
| normalization (36,063) | primary (922) | raw sequence (115) |
| mirna transcript (26,715) | mirna sequence data (833) | other (106) |
| mirna sequence (26,715) | assembly (815) | molecular data (106) |
| scan (24,865) | nucleic acid structural information (770) | recording acoustical (102) |

| | | |
|---|---|---|
| genomic sequence data (24,293) | profile (754) | non coding rna profiling by genome tiling array (98) |
| gene and protein information (24,292) | probe logratios (754) | case set (97) |
| image (11,778) | probe calls (754) | cohort (84) |
| supplementary material (11,374) | gene logratios (754) | analysis results (84) |
| expression profiling by high throughput sequencing (10,162) | gene calls (754) | genome variation profiling by high throughput sequencing (70) |
| protein coding (10,054) | two columns (691) | tabular digital data (69) |
| processed data matrix (9,741) | profiles (666) | raw data matrix (61) |
| protein (8,739) | expression profiling by genome tiling array (620) | family (56) |
| r object (7,026) | mageml (597) | longitudinal (54) |
| gene sequence data (6,162) | genome variation profiling by array (587) | |
| nmr results (6,157) | genome variation profiling by snp array (555) | |

# Appendix C. Chemistry: All data types that received at least one data citation

(data type, number of total data citation)

| | | |
|---|---|---|
| crystal structure (754,913) | still images or photos (63) | pka determination data (1) |
| crystallographic data (490,252) | molecule characterization (38) | pictures (1) |
| molecular structure (91,870) | envelope stored as signed bytes (7) | nmr titration data (1) |
| crystallographic information (84,687) | structural model (4) | nmr data (1) |
| bacterial carbohydrate structure (4,298) | micro electron diffraction (3) | metadata (1) |
| spectral data (3,720) | xfel diffraction (1) | gigadb dataset (1) |
| crystallographic structure (3,008) | x ray diffraction images (1) | diffraction images (1) |
| dataset (2,410) | x ray diffraction data (1) | crystal x ray structure (1) |
| molecular data (954) | structures (1) | chloride binding data (1) |
| molecule (647) | structure fragments (1) | bacterial carbohydrate structures (1) |
| image stored as reals (379) | raw crystallography data (1) | anion transport data (1) |
| x ray diffraction (255) | primary data nmr mass ir raman xray tlc (1) | analytical data (1) |

## Appendix D. Computing: All data types that received at least one data citation

(data type, number of total data citation)

| | | |
|---|---|---|
| software (18,246) | survey and census data (1) | GIS data (1) |
| code (1,278) | spreadsheet (1) | earth and environmental data (1) |
| model (416) | simulation MATLAB code (1) | diagrams (1) |
| dataset (3) | raw data (1) | dataset used in the paper (1) |
| raw experimental data (2) | open source coding and tools (1) | chemistry data (1) |
| other (2) | network data extracted from social media (1) | |
| database (2) | life science database (1) | |

## Appendix E. Earth sciences: All data types that received at least one data citation

(data type, number of total data citation)

| | | |
|---|---|---|
| dataset (32,975) | seismic shottimes mcs (73) | physicalproperties sediment (4) |
| interactive resource (22,264) | chemistry fluid (72) | dopplervelocity (4) |
| GPS dataset (13,080) | seismic reflection mcs (68) | digital map data (4) |
| geoscientific information (9,108) | terrestrial lidar point cloud (63) | technicalreport (3) |
| GPS collection (5,741) | gravity anomaly freeair (60) | seismic shottimes scs (3) |
| text (4,033) | currentmeasurement (60) | sample rock ancillary (3) |
| navigation primary (3,691) | seismic ancillary mcs (53) | physicalproperties sediment ancillary (3) |
| protein sequence data (2,803) | seismic active subbottom (40) | oceanographic data (3) |
| digital (2,699) | seismic segyhistory mcs (38) | interpretation geologic (3) |
| image (1,063) | radiation (38) | chemistry sediment (3) |
| observational data (927) | gis vector data (33) | biology species abundance (3) |
| bathymetry singlebeam (841) | bathymetry swath ancillary (31) | biology microbiology (3) |
| gravity field (738) | seismic navigation (28) | application pdf (3) |
| magnetic field (683) | backscatter optical (27) | visualization (2) |

| | | |
|---|---|---|
| temperature (662) | geoid ondulation given on a grid (26) | turbulence (2) |
| bathymetry swath (643) | photograph webgallery (22) | spatial characteristics (2) |
| meteorological (534) | magnetic anomaly igrf (21) | scanned map (2) |
| imagedigital (417) | photograph (20) | satellite imagery (2) |
| bathymetry (405) | dkrz series technical report (19) | rainfall patterns (2) |
| conductivity (390) | bathymetry phase (19) | population (2) |
| navigation (335) | turbidity (18) | physicalproperties rock (2) |
| sidescan (280) | visualization googleearth (15) | particleflux (2) |
| backscatter acoustic (263) | aerial or satellite imagery (15) | |
| salinity (254) | text tab separated values (14) | |
| software (253) | seismic shottimesstatus (13) | |
| radiation visible (234) | photograph mosaic (9) | |
| radiation infrared (221) | geoid undulation given on a grid (9) | |
| fluorescence (210) | digital terrain model (8) | |
| pressure (184) | seismic ancillary scs (7) | |
| ctd ancillary (179) | transmissivity (6) | |
| seismic reflection scs (147) | digital map (6) | |
| mapdigital (147) | biology species list (6) | |

| | | |
|---|---|---|
| iceconcentration (112) | chemistry fluid<br>electrochemistry (5) | |
| soundvelocity (92) | tabledigital (4) | |

Appendix F. Engineering: All data types that received at least one data citation

(data type, number of total data citation)

| | | |
|---|---|---|
| test data (3,749) | qcm data (1) | excel spreadsheet (1) |
| dataset (4) | microscopy images (1) | datasets containing results of materials testing and accompanying information (1) |
| GIS vector data (2) | fluorescence intensity data (1) | |

Appendix G. Mathematical sciences: All data types that received at least one data citation

(data type, number of total data citation)

| | |
|---|---|
| software (8,155) | geoid undulation given on a grid (35) |
| matrix (1,640) | |

# Appendix H. Technology: All data types that received at least one data citation

(data type, number of total data citation)

| | | |
|---|---|---|
| dataset (137,375) | application x rar (23) | data (4) |
| fileset (33,304) | image x ms bmp (18) | audio x aiff (4) |
| image tiff (14,558) | excel (18) | application x bzip2 (4) |
| image (12,591) | mixed (17) | video 3gpp (3) |
| application MS Word (11,482) | raw data (15) | thesis doctoral (3) |
| software (8,176) | image gif (15) | results (3) |
| application pdf (6,626) | figure data (13) | provenance files and benchmark data (3) |
| application vnd MS Excel (3,495) | composite document file v2 document corrupt can't expand summary info (13) | performance results (3) |
| tools (2,428) | composite document file v2 document no summary info (12) | nnmr spectroscopic (3) |
| text plain (686) | text x PERL (10) | linked data endpoint access logs (3) |
| application octet stream (664) | microsoft excel (9) | image svg xml (3) |
| video quicktime (528) | table (7) | GIS vector data (3) |

|  |  |  |
|---|---|---|
| application postscript (527) | image x coreldraw (7) | excel spreadsheets in zipped format (3) |
| video x msvideo (463) | application x 7z compressed (7) | data from publication (3) |
| image jpeg (421) | quantitative (6) | base (3) |
| video mp4 (249) | excel spreadsheet (6) | audio files (3) |
| application zip (246) | excel file (6) | newscutting (2) |
| audiovisual (138) | audio mpeg (6) | model (2) |
| image png (119) | text x tex (5) | MATLAB (2) |
| application vnd MS Powerpoint (116) | text x fortran (5) | images txt files (2) |
| video x ms asf (106) | text x c (5) | geospatial (2) |
| video mpeg (94) | spreadsheets (5) | figures (2) |
| text html (93) | source code (5) | fig (2) |
| video protocol (78) | gle (5) | excel data and images (2) |
| text rtf (71) | dat (5) | eps (2) |
| database (64) | csv (5) | dataset for figure (2) |
| compact model (46) | code (5) | data series (2) |
| spreadsheet (34) | video (4) | composite document file v2 document corrupt cannot read summary info (2) |
| supplementary material (32) | still images or photos (4) | audiovisual data (2) |
| audio x wav (32) | sound (4) | application x tar (2) |

| | | |
|---|---|---|
| type of data field content (29) | origin files (4) | application x shockwave flash (2) |
| application xml (29) | moving image (4) | application ogg (2) |
| application x gzip (28) | image vnd Adobe Photoshop (4) | |
| text (23) | experimental data (4) | |

CURRICULUM VITAE

**Hyoungjoo Park**

Doctoral Candidate
School of Information Studies, University of Wisconsin - Milwaukee

## EDUCATION

University of Wisconsin - Milwaukee, School of Information Studies, Milwaukee, WI, USA

- PhD candidate
  Committee Members: Dr. Dietmar Wolfram (Chair), Dr, Richard P. Smiraglia, Dr. Jacques du Plessis, Dr, Margaret E.I. Kipp, & Dr. Catherine Blake (external member at University of Illinois at Urbana-Champaign)

Syracuse University, School of Information Studies, Syracuse, NY, USA

- Master of Science in Information Management

SungKyunKwan University (SKKU), Seoul, Korea

- Bachelor of Library and Information Science
- Bachelor of Business Administration

## JOURNAL PUBLICATIONS

(* means I am the corresponding author)

1. **Park, H**., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and data re-use are more common than formal data citation in biomedical fields. *Journal of Association for Information Science and Technology, (69)*11, 1346-1354.
2. **Park, H**.*, & Wolfram, D. (2017). Research data sharing and re-use: Implications for data citation practice. *Scientometrics, 111*(1), 443-461.
3. Smiraglia, R.P. & **Park, H.** (2017). Ontological data-sharing of open government data for data curation. *Canadian Journal of Information and Library Science, 41*(4), 285-307
4. Cai, X., Castillo, M., Graf, A., Hassan, M., **Park, H.** & Smiraglia, R.P. (2016). Knowledge organization and the 2015 UDC seminar: An Editorial. *Knowledge Organization, 43*(6), 395-402.
5. Beak, J., Choi, I., Lee, S., **Park, H.**, Ridenour, L., & Smiraglia, R.P. (2014). Knowledge organization and the 2013 UDC seminar: An Editorial. *Knowledge Organization, 42*(3), 191-194.

## CONFERENCE PAPER, POSTER, INVITED TALK

1. **Park, H.*** (accepted). Implications of data sharing on formal data citation in biomedical fields. *iConference 2019*.
2. **Park, H.*** & Kipp, M.E.I. (2018). Library linked data models: Multinational library data in the semantic web. *ASIS&T Workshop: Big Metadata Analytics.* Vancouver, Canada.
3. **Park, H.**, & Wolfram, D. (2018). Research data sharing and re-use: Practical implications for data citation practice that benefit researchers. *Southern California Clinical and Translational Science Institute*, *University of Southern California*, CA. (invited speaker)
4. **Park, H.*** (2018). The impact of research data sharing and re-use on data citation in STEM fields. *Association for Library and Information Science Education (ALISE) Jean Tague Sutcliffe Doctoral Student Research Poster Competition,* ALISE 2018 Annual Conference, Denver, CO.
5. **Park, H.*** (2017). The impact of research data sharing and re-use on data citation in STEM fields. *ASIS&T Doctoral Seminar on Research and Career Development*, Washington, DC.
6. **Park, H**., You, S., & Wolfram, D. (2017). Is Informal data citation for data sharing and re-use more common than formal data citation? *In ASIS&T 2017 Proceedings.*
7. **Park, H.**, & Wolfram, D. (2017). Formalised data citation practices would encourage authors to make their data available for reuse. *LSE Impact Blog of the London School of Economics and Political Science*.
8. Lee, T., **Park, H.**, Lee, S. & Choi, I. (2017). Practice of multilingual supports: A pilot study of the top 25 circulation public libraries in the ALA. *In iConference 2017 Proceedings.*
9. Lee, T., **Park, H.**, Lee, S. & Choi, I. (2017). Public libraries' multilingual supports: A preliminary framework in 19 public libraries in United States. *ALISE 2017 Annual Conference*. January 17–20, 2017, Atlanta, GA.
10. **Park, H.*** (2016). Data citation practices in scientific research communities: The Data Citation Index of Web of Science. *Research Data Access and Preservation (RDAP) Summit 2016*. May 4-5, 2016. Atlanta, GA.
11. Smiraglia, R.P. & **Park, H.** (2016). Using Korean open government data for data-curation and data integration. *In Dublin Core and Metadata Applications (DCMI) 2016 Proceedings.*
12. **Park, H.*** (2015). From industry to scholarly communication: Biometric literature over time. *In iConference 2015 Proceedings.*
13. **Park, H.***, & Kipp, M.E.I. (2015). Evaluation of mappings from MARC to Linked Data. *Advances in Classification Research Online*. *(ASIS&T SIG/CR Paper)*
14. **Park, H.*** (2015). From Industry to Scholarly Communication: Biometric Literature Over time. *iConference 2015 Annual Conference.* March 24-27, 2015. Newport Beach. CA.
15. **Park, H.***, & Smiraglia, R. P. (2014). Mapping Korean Open Government Cultural Heritage Data with CIDOC CRM for Data Curation: A Case Study. *CRMEX 2014*.
16. **Park, H.*** (2014). Integrated Geographic Data Visualization with Open Government Data: Seoul Metropolitan Government of Korea. *The 9th International Digital Curation Conference (IDCC),* February 24-27, 2014. San Francisco, CA.

17. **Park, H.***, & Smiraglia, R.P. (2014). *Enhancing Data Curation of Cultural Heritage for Information Sharing: A Case Study using Open Government Data*. Communications in Computer and Information Science. 478, 95-106. Switzerland: Springer International Publishing.
18. **Park, H**.*, & Lee, S. (2014). Cultural analysis for current situations of Asian students' school adjustment - case study of Korean students at the University of Wisconsin – Milwaukee. *Trans-Asia Graduate Student Conference*. April 5–6, 2014, Madison, WI, USA
19. **Park, H.**, & Smiraglia, R.P. (2014). Enhancing Data Curation of Cultural Heritage for Information Sharing: A Case Study using Open Government Data. *Knowledge Organization Research Group. University of Wisconsin – Milwaukee*. November 13, 2014.
20. Ridenour, L., & **Park, H**.* (2014). Visualizing Analysis of Historical Patterns with Word Collocation: ISKO Proceedings Titles from 1992 to 2012. *The 13th International Society for Knowledge Organization (ISKO) Conference*. May 19–22, 2014. Krakow, Poland
21. **Park, H.** (2009). Personal Information Management. *Korean Agency for Standards and Technology (KATS),* Seoul, Korea.

## RESESARCH EXPERIENCE

1. Data Citation Index vs. Event Data Index (November 2018 ~ April 2019)
    - Funded by National Science Foundation – Merging Science and Cybereinfrastructure Pathways: The Whole Tale (Whole Tale)
    - Funded by Alfred P. Slogan Foundation – United States region of Research Data Alliance (RDA/US) Data Share Fellowship
    - role: data fellow (Whole Tale – RDA/US fellowship)
        - comparative analysis of Data Citation Index and Event Data Index
        - writing project report
        - presenting project in 2019 Research Data Alliance (RDA)
2. Open Peer Review (Professor Dietmar Wolfram) (Summer 2018 ~ December 2018)
    - Funded by UWM Internal Grant – Research Growth Initiative.
    - role: research assistant
        - Models for open peer review journals
3. Knowledge Organization (Professor Richard Smiraglia) (Fall 2013 ~ Summer 2015)
    - role: research assistant
        - Domain analysis for knowledge organization
        - Ontological data sharing for knowledge organization
4. Metadata quality evaluation - Linked data (Associate Professor Margaret E.I.Kipp) (Fall 2013, Fall 2014 ~ Summer 2015)
    - role: research assistant
        - Metadata quality evaluation for linked data
        - Metadata user study on linked data

## TEACHING EXPERIENCE

**School of Information Studies, University of Wisconsin – Milwaukee**

**Lead Instructor (Undergraduate, Face-to-face courses)**

1. Lead Instructor, INFOST 230 Organization of Knowledge (Spring 2018)
2. Lead Instructor, INFOST 325 Information Security I (Fall 2017)
3. Lead Instructor, INFOST 230 Organization of Knowledge (Fall 2017)
4. Lead Instructor, INFOST 230 Organization of Knowledge (Spring 2017)
5. Lead Instructor, INFOST 230 Organization of Knowledge (Fall 2016)

**Teaching Assistant (Graduate, Online courses)**

1. Teaching Assistant, INFOST 582 Introduction to Data Science (Spring 2016)
2. Teaching Assistant, INFOST 511 Organization of Information (Spring 2016)
3. Teaching Assistant, INFOST 511 Organization of Information -section I (Fall 2015)
4. Teaching Assistant, INFOST 511 Organization of Information – section II (Fall 2015)

**GangNeung Continuing Education Center,** Korea

**Lead Instructor (School Media Librarian, Face-to-face course)**

1. Lead Instructor, Introduction to Library and Information Science

**LG CNS Academy of Information Technology, LG CNS Headquarters**, Seoul, Korea

1. Instructor: various (Winter, 2010)

**School of Information Studies, Syracuse University**

**Teaching Assistant (Graduate, Face-to-face courses)**

1. Teaching Assistant, IST 659 Data Administration Concepts and Database Management (Summer 2004)
2. Teaching Assistant, IST 758 Designing Web-based Database Systems (Summer 2003)
3. Teaching Assistant, IST 637 Database Management for Library Services (Summer 2003)

## AWARDS/ FELLOWSHIPS

1. Early Career Fellowship, Whole Tale – US region of the Research Data Alliance (WT-RDA/US). (2018) ($5,000 + travel support)
2. 2018 Eugene Garfield Doctoral Dissertation Fellowship, Beta Phi Mu International Honor Society (2018) ($3,000)

3. Travel Awards, Graduate School, Graduate School, University of Wisconsin – Milwaukee (2018) ($600)
4. SOIS PhD Scholarship, School of Information Studies, University of Wisconsin – Milwaukee. (Spring, 2018). ($2,011.41)
5. Research Assistantship, full tuition remission & stipend. Internal Grant - Research Growth Initiative (RGI), University of Wisconsin-Milwaukee (Summer 2018 ~ current).
6. SOIS PhD Scholarship, School of Information Studies, University of Wisconsin – Milwaukee. (Fall 2017) ($2,011.41)
7. Travel Awards, Graduate School, Graduate School, University of Wisconsin – Milwaukee (January 2018) ($500)
8. Chancellor's Graduate Student Awards, University of Wisconsin -Milwaukee (October 2017) ($1,050)
9. Travel Awards, School of Information Studies, University of Wisconsin -Milwaukee (October 2017) ($900)
10. ASIS&T, invited for ASIS&T Doctoral colloquium (registration fee waiver), Association for Information Science and Technology (October 2017)
11. Chancellor's Graduate Student Awards, University of Wisconsin -Milwaukee (Fall 2016) ($1,000)
12. Travel Awards, School of Information Studies, University of Wisconsin -Milwaukee (January 2017) ($1,000)
13. Travel Awards, School of Information Studies, University of Wisconsin -Milwaukee (May 2016) ($500)
14. SOIS PhD Scholarship, SOIS, University of Wisconsin -Milwaukee (Spring 2015) ($580)
15. Travel Awards, School of Information Studies, University of Wisconsin -Milwaukee (December 2014) ($1,500)
16. Doctoral Student Publication Award, School of Information Studies, University of Wisconsin -Milwaukee (December 2014) ($300)
17. SOIS PhD Scholarship, School of Information Studies, University of Wisconsin -Milwaukee (Fall 2014) ($583)
18. Chancellor's Graduate Student Awards, University of Wisconsin -Milwaukee (Fall 2014) ($1,000)
19. IOrg Conference Support, Information Organization Research Group, School of Information Studies, University of Wisconsin -Milwaukee (Summer, 2014) ($500)
20. Travel Awards, School of Information Studies, University of Wisconsin -Milwaukee (Spring 2014) ($1,500)
21. SOIS PhD Scholarship, School of Information Studies, University of Wisconsin -Milwaukee (Spring 2014) ($605)
22. Chancellor's Graduate Student Awards, University of Wisconsin -Milwaukee (Fall 2013) ($1,500)
23. SOIS PhD Scholarship, School of Information Studies, University of Wisconsin -Milwaukee

(Fall 2013) ($605)

24. Research & Teaching Assistantship, full tuition remission & stipend. School of Information Studies, University of Wisconsin -Milwaukee (Fall 2013 ~ Spring 2018)
25. LG CNS chief executive officer (CEO)'s best R&D award of the year, LG CNS Headquarters, Seoul, Korea (2010)
26. Winner of the year in Recognition of Excellence in analyzing external R&D ideas, LG CNS Institute of Information and Technology, LG CNS Headquarters, Seoul, Korea (2007)
27. Full tuition remission for international exchange student program, SungKyunKwan University (SKKU), Seoul, Korea (2001-2002)

## HONOR SOCIETY

Phi Beta Delta, Honor Society for International Scholars, Syracuse University (joined in 2004)
Golden Key International Honor Society (joined in 2014)

## CERTIFICATES

Certificate of Librarianship, Korean Library Association (KLA)

## SERVICE

- Committee
    - PhD student representative, Academic Planning Committee, School of Information Studies, University of Wisconsin – Milwaukee (Fall 2015 ~ Spring 2017)
    - Editor, School of Information Studies, University of Wisconsin – Milwaukee (2015)
    - Executive Committee Officer, Doctoral Student Organization, School of Information Studies, University of Wisconsin – Milwaukee (Fall 2014 ~ Spring 2016)
    - Young Leaders. LG CNS Co., Ltd. Seoul, Korea
- Journal Reviewer
    - Transactions on Emerging Topics in Computing (ad hoc)
    - Knowledge Organization (ad hoc)

## PROFESSIONAL MEMBERSHIP

Association for Education in Library and Information Science (ALISE)
Association for Information Science and Technology (ASIS&T)