MEASURING AND ANTICIPATING THE IMPACT OF DATA REUSE

by

Kathleen Marie Fear

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2013

Doctoral Committee:
Professor Elizabeth Yakel, Chair
Assistant Professor Eytan Adar
Professor George C. Alter
Professor Margaret L. Hedstrom

ACKNOWLEDGEMENTS

I would like to thank my dissertation chair, Elizabeth Yakel, for her guidance and support throughout my time in the doctoral program. Working with Beth on my master's thesis project inspired me to continue on to the doctoral program, and her ongoing encouragement kept me moving forward despite my missteps and stumbles. I am also especially grateful to Margaret Hedstrom, who brought me into the NSF IGERT Open Data program (National Science Foundation Grant No. 0903629), and in doing so played a major role in shaping my research and professional career. George Alter provided ready advice and guidance throughout this dissertation study that substantially shaped its intellectual direction, and I am also grateful for the generous access he granted me to the usage data and ICPSR study metadata that forms the core of my study. I would also like to thank Eytan Adar for his guidance and suggestions throughout the process of this dissertation. I could not have asked for a better committee: you continually challenged me to produce my best work, and I feel privileged to have had the chance to work closely with you all.

This dissertation would not have been possible without generous access to data given to me by several sources. The bibliographic data I used came from the ICPSR Bibliography of Data-Related Literature, which is produced and maintained by Elizabeth Moss. I would also like to thank Ixchel Faniel, Elizabeth Yakel, Adam Kriesberg and Morgan Daniels for their work on the Dissemination Information Packages for Information Reuse

(DIPIR) project, Institute for Museum and Library Services, Grant # LG-06-10-0140-10 and for granting me access to the interview data from data reusers.

I would also like to thank all those who provided support of all kinds throughout this project. Veronica Falandino, Jen Todd, Sue Schuon, Karen Woollams and Lai Tutt were always willing and ready to answer my questions about the doctoral program, from travel funding to dissertation deadlines. I want to thank Dharma Akmon for reading and reviewing endless revisions of my proposal and then dissertation, and Matt Burton and Morgan Daniels for their valuable feedback as I worked through my dissertation proposal. Throughout my time at SI, I had the opportunity to work with excellent collaborators, all of whom contributed to my work and intellectual growth through their kindness, advice and support, especially Devan Donaldson, Paul Conway, and Kai Zheng. I am grateful to Mary Ann Mavrinac and Katie Clark at the University of Rochester for their patience and flexibility as I have worked to complete my dissertation and begin my career.

I am thankful to my parents for the sacrifices they made to send me along this path. Finally, I am grateful to my husband, Ryan, for his unwavering support and confidence in me, and to my sister, for always knowing how to lift my spirits.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

Introduction

1.1 Background and problem statement

In 2013, an economics graduate student identified an error made by high-profile researchers in a 2010 paper that had been an important influence on public policy (Herndon, Ash, & Pollin, 2013). In 2010, after several years of planning and work to coordinate and share data, a collaborative group of Alzheimer's researchers began making breakthroughs on the detection and early diagnosis of that disease (Mueller et al., 2005). Between 2007 and 2010, scientists in genomics produced 1,150 new papers from data they did not collect themselves (Piwowar, Vision, & Whitlock, 2011). These successes in research were all made possible through data sharing: directly, in the case of the economics paper and through collaborative data production, in the case of the Alzheimer's Disease Neuroimaging Initiative; and through contribution of data to a repository, the Gene Expression Omnibus. Scholars, funding agencies and public policy makers increasingly recognize sharing data for others to reuse as an important part of scholarship.

Sharing data goes hand-in-hand with preserving them. There are numerous potential benefits to preserving data (Beagrie, Chruszcz, & Lavoie, 2008). Preserving and sharing data can increase the return on investment in research by ensuring the persistence of

unique observations, for example observational data collected on atmospheric phenomena, or data collected on threatened populations or species. Shared data opens the door to further research that builds on an original set of findings or supports an innovative repurposing of data. If researchers beyond the original project can squeeze more value out of data, they can produce more science and more knowledge from that initial investment than would otherwise have been possible. Enabling data reuse through curation and long-term management of data though, is a costly proposition in and of itself. Furthermore, scientists lack incentives to share data. One reason for this is that unlike other forms of scholarship, there are not yet metrics for measuring and rewarding the contribution that data sharing makes.

In scholarship, valuable contributions of knowledge are those that ripple outward, and so research value is often identified via its impact. High impact research is research that influences other work, typically quantified (at least in part) by the number of times a particular piece of research is referenced elsewhere. In other words, how often do other people point to that work as a direct influence on their own work, as part of a foundation of prior work, or as a starting point for new ideas? Metrics, such as the h-index (Hirsch, 2005) and g-index (Egghe, 2006) use the number of highly cited papers in an author's body of work as a measure of the impact his or her research has had. This approach is not without problems, particularly in that authors may cite a previous article as an example of poorly done work or a mistaken interpretation, but in general, counting metrics like the h- and g-index represent a key conceptualization of value in scholarly work: the best-of-the-best journal articles are perceived to be those that seed many others.

Impact metrics identify highly achieving scientists, and as a result, they are often used in the processes of hiring faculty and granting tenure, either formally or informally. On one hand, these measures provide an incentive to publish widely and share results through the scientific literature, but the importance placed on them sets up a series of "perverse incentives" (Kapeller, 2010) that can encourage scientists to focus on the production of citable articles. The dichotomy between work that is rewarded and work that is not can be problematic. Work that is not directly rewarded may be seen as secondary, but it is not necessarily unimportant or valueless. Published journal articles are the primary rewarded product of science, but scientists engage in a wide range of other activities in service of the production of papers, among them data management.

Sharing data opens up a new area of scholarly contribution: use of a dataset by scientists other than the data producers, or in other words, reuse. Data sharing contributions, if they are acknowledged at all, are not rewarded to the same extent that contributions in the form of journal articles are (McDade et al., 2011). While the initial work of managing data to produce a paper is rewarded through the citations that accrue to the paper, there is no mechanism to acknowledge use of the data beyond the original producers, or in other words, to acknowledge data reuse. This study focuses on the question of measuring reuse impact: how can the contribution scientists make when they share their data be recognized and quantified?

Some datasets are well known to foster very high levels of reuse. The National Longitudinal Study of Adolescent Health (Add Health), for example, has enough users to support an annual conference,[1] and it has produced well over 4,000 journal articles, conference papers, dissertations, and other publications. By standard counting metrics, which reward raw performance in the form of citations, Add Health and datasets like it are high impact and thus valuable. But these metrics also embed a very particular perspective on value: the more people that specifically refer back to a product of scholarship, the more valuable it is.

Following the model of publication impact, one might suppose that the most valuable datasets are those that are reused many times: datasets from which a number of different questions can be asked and answered. But it may be the case that unlike knowledge embedded in papers, the influence of data is not directly identifiable through the number of citations to the dataset. If, for example, a researcher produces a dataset and from those data writes an important paper that sparks a huge amount of subsequent work, it is likely that this subsequent work will cite the paper, not the data. Indeed, sharing data is associated with an increase in citations to the paper describing the data (Piwowar, Day, & Fridsma, 2007). Citation counts for that paper will be high, but there may not be many (or even any) new citations to the data even though the data played a significant role in sparking new lines of inquiry and arguably have significant preservation value. How can we take into account such downstream impacts when considering the value of a dataset?

[1] http://www.cpc.unc.edu/projects/addhealth/news/2012-add-health-users-conference

Data sharing incurs additional responsibilities for data management not only on the part of data producers but also for repositories that curate and manage data over the long term. Some datasets are reused over and over again by many researchers; others, despite undergoing the same careful curation processes, may never be touched again. Repositories need to make forward-looking decisions about how much of their resources to devote to curating any given dataset, which raises the question of how to gauge whether a dataset is likely to be reused. This study addresses the issue of anticipating impact: are there indicators of whether a dataset is likely to be reused in the future that can be assessed at the time data are produced or deposited to a repository?

Data sharing opens a new opportunity within the scholarly reward system. When scientists share data, they make an important scholarly contribution, but there is currently no accepted way to measure and recognize that contribution. This study attempts to bridge this gap by proposing data reuse impact metrics and exploring the characteristics of data that influence their reuse impact.

1.2 Research questions and design overview

This study is a quantitative exploration of data reuse in three parts: 1) identifying citation patterns among data reusers; 2) developing and demonstrating a suite of data reuse impact metrics; and 3) exploring factors that influence whether or not a dataset is reused, as well as what impact its reuse has.

The guiding research question is: *What is the scholarly impact of data reuse? How can stakeholders anticipate the impact the data they fund, create or curate will have?*

The specific research questions are:

- How and why do social scientists cite data?
- Which datasets held by ICPSR are high-impact according to different measures of reuse impact?
- What characteristics of data predict whether they will be reused?

To address the first question, I examined publications related to a sample of datasets held at The Inter-university Consortium for Political and Social Research (ICPSR). In the second part of the study, I used the citation information I compiled in the first part to demonstrate four proposed metrics of reuse impact. These metrics incorporate existing methods of accounting for research impact and also go beyond counting citations or other indicators of usage to account for the broader network of reuse around data. Using additional bibliometric measures, these metrics quantify the different ways in which a dataset can have impact beyond direct usage of the data. I addressed the final question in two parts. First, I identified a set of factors that may influence whether data is reused or not drawn from the substantial literature on data reuse in a variety of disciplines. I analyzed interviews with data reusers in the social sciences to validate and extend the list of characteristics drawn from the literature. Using the finalized list of factors, I produced a series of regression models exploring the relationships between these factors and whether or not a dataset is reused, as well as what impact that reuse has had.

1.3 Key findings

I identified key findings from each section of the study, and together, they address the overarching research question. Examining data reuse in the social science literature

showed that secondary users of data regularly cite data producers' publications, rather than citing data directly or crediting the data provider (in this case, ICPSR). Where they do provide citations to the data provider, they typically do so in addition to citing the data producer: the data provider citation does not replace the citation to the data producer. This finding suggests that data reusers distinguish between the contributions producers make when they create data and when they share it: in essence, data reusers use citations to credit both actions. In the next section, I explored reuse impact. The four measures of reuse impact I developed highlighted different aspects of impact for data; no datasets were high-impact across the board, and few were consistently low-impact. The three metrics based on citations were especially divergent, suggesting that data can have an impact in multiple and varying ways. Finally, Chapter 6 showed that two characteristics of data are particularly related to whether the data are reused or not: the size of the data and how actively used they are.

Together, these findings indicate that sharing data makes a contribution to scholarship above and beyond the initial contribution a scientist makes when she creates data and publishes from them. In the sample of datasets I studied, Big Data are not the only important data; active data—those whose producers write multiple articles, in particular—were also heavily reused. Social scientists' citation practices credit data sharing separately from data production; their reuse of data demonstrates the multiple ways in which shared data makes an impact.

1.4 Data and data reuse in the social sciences

Prior efforts to analyze data reuse have been hindered by the difficulty in gathering the corpus of data that is required for bibliometric analysis; citations to data tend to be unevenly present and idiosyncratic where they do exist. The social sciences represent a unique opportunity in this sense. Weber and Chao (2011) found that reuse of quantitative social science data held by repositories was both well documented and relatively consistent in citation format; further, reuse of quantitative data from repositories represents the most common type of data reuse in these fields.

The social sciences have a long tradition of sharing data and conducting secondary analyses or reanalyses using existing data. This has to do in large part with the nature of social research, which often requires large amounts of unique data as well as data collected over time. As a result, many social scientists have historically relied on data collected by others, both in the form of large-scale data collection efforts like the Census and the General Social Survey as well as smaller-scale efforts by individual researchers or small groups. Much social science data, though, is small data, and that is the data that is most at risk (King, 2011). Large-scale data collection efforts often come with funding intact for their curation and ongoing preservation. Data produced by individuals, though, may not have this same kind of support.

The small data that underlie so many publications are crucial to share and maintain in order to ensure the replicability of social science research (King, 2011). Reuse of smaller datasets in the social sciences is not well studied, and it is these data that are the special

focus of this study. Developing impact metrics and an understanding of what makes data likely to be reused is especially important for such data. These data are too often 'dark' and are likely to remain dark if their value is not recognized. This study shines a light on the ways in which data, large or small, can be valuable, and it provides some guidance for strategizing curation and preservation of small data.

1.5 Data sources

This study examined the reuse of three cohorts of social science datasets made publically available in 2000, 2001, and 2002 by The Inter-university Consortium for Political and Social Research (ICPSR). These datasets comprise a range of disciplines and data types and are a mix of studies produced by individuals, groups of researchers and governmental agencies and organizations. Together, these datasets represent the broad variation in social science data, with an emphasis on small-scale data collection efforts.

ICPSR, a data repository housed at the University of Michigan, is a major source of social science data, with holdings containing more than 65,000 individual datasets from about 8,000 studies. The repository is host to several major social science data collections, including the US Census, the General Social Survey, the American National Election Study and the American Community Survey, as well as studies conducted by individual researchers. ICPSR's goal is not only to preserve social science data over the long term, but to also make the data accessible for reuse. ICPSR's data holdings cover a range of disciplines within the social sciences, and any researcher affiliated with any ICPSR member institution has access to all datasets, regardless of his or her own disciplinary affiliation. The diversity in data at ICPSR potentially enables

interdisciplinary reuse, especially with the assistance of ICPSR's several tools for data discovery that allow searching and browsing across the full range of the repository. ICPSR has a well-developed data curation strategy aimed at making data maximally available for reuse, but the datasets the repository houses inherently differ in their nature, the contexts in which the data were produced, and other factors.

Much of ICPSR's data is publically available for download, and individuals associated with member organizations have broader access to additional data. ICPSR provides boilerplate citation text as a standard part of data downloads, which makes it easy for authors to drop that information directly into their manuscripts. Furthermore, the repository devotes significant resources to tracking reuse of its holdings and compiling that information into a citation database. The ICPSR Bibliography of Data-Related Literature[2] enables the repository to present information alongside datasets about the papers that have used that data.

The Bibliography is also a valuable resource for research in and of itself, and it is a central source of data for my study. The ICPSR Bibliography of Data-Related Literature is comprised of citations from journal articles, books, book sections, conference papers and gray literature like reports and unpublished theses to data at ICPSR. I used the Bibliography to track reuse of ICPSR's datasets in the social science literature. Specifically, I mined the Bibliography for the citation related to dataset in my sample and

[2] http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/

used that information to identify patterns in data citation practices as well as to compile citation statistics, which formed the basis for computing data reuse impact metrics.

I derived further information about these datasets from a number of sources. ICPSR's metadata records encode detailed study-level information about the datasets, along with the names and affiliations of the data producers. Using Scopus, I collected information about data producers, including their co-authorship networks and their publication networks. The same source provides further information about the data-related publications in the ICPSR Bibliography, including papers that cite those documents. Finally, ICPSR provided usage records of its datasets, specifically download statistics for the datasets in my sample. Each data source and how I derived data from it is described in more detail in Chapter 3, Methods.

1.6 Definitions

This dissertation uses several specialized terms, as well as common terms used in specialized ways. This section introduces important terms and their meanings.

Dataset: For this dissertation, the term dataset refers to the organized collection of data produced in a single study (i.e. not data that is part of a repeated series of data collections) by an individual or group of data producers or by an organization. The dataset may be spread over one or more data files. I use 'dataset' generally to refer to a specific dataset (e.g. "The dataset has been downloaded 500 times"), and 'data' to refer to multiple datasets or data in general (e.g., "Social science data are important").

Study: A study is comprised of a dataset (as described above) plus a codebook produced by ICPSR along with any supplemental documentation, such as setup files, questionnaires, etc. ID numbers at ICPSR are assigned at the study level. For example,

the Six-State Survey of Elderly Dual Enrollees in Medicare and Medicaid, 1999[3] contains a single datasets along with a codebook. This collection of materials forms a single coherent study, with ICPSR ID number 3240. Descriptive metadata is provided at the study level.

Data producer: I use 'data producer' to refer to the individual, group of individuals, or organization that is responsible for creating the data in a given dataset. The data producer can create data either through a research project or as a byproduct of other work (as in the case of administrative data). The term 'producer' does not include a repository that takes responsibility for curating and disseminating a dataset.

Data provider: The term 'data provider' refers to an entity that provides access to data. This can be a repository, such as ICPSR, or the data producer herself.

Administrative data: Administrative data are data derived from information collected in the course of carrying out an organizational function. An example of administrative data is criminal arrest records, which are collected in the course of law enforcement activities. These data can be used for research, but they are not created with a specific research question in mind.

Data reuse: Data reuse is use of data one which did not collect oneself, for example to answer a new question from existing data (Zimmerman, 2008), to combine with other existing or newly collected data, or to reproduce or replicate the results of a prior study (King, 1995). Data producers repurposing their own data is not considered data reuse for the purposes of this study. In the social sciences, the terms secondary analysis (Gleit & Graham, 1989; Hinds, Vogel, & Clarke-Steffen, 1997) or reanalysis (King, 2003; Weber & Chao, 2011) are often used synonymously with my definition of data reuse.

Data-related publication: For the purposes of this study, I refer to publications that involve data from ICPSR as 'data-related publications.' The terminology is borrowed from ICPSR, which compiles citations from publications that use ICPSR data in their bibliography of data-related literature. Both producers and reusers of a dataset create data-related publications. Data-related publications can include published items like journal articles, editorials, letters to the editor, literature reviews as well as unpublished reports and other gray literature.

Secondary publication: Secondary publications are the subset of data-related publications that use a dataset but are not written by the producers of that dataset. More specifically, secondary publications are publications that reuse data held by ICPSR. This term is used synonymously with 'reuse publication.' I differentiate between secondary publications, which reuse data, and other publications not written by the data producer that reference but do not reuse data, such as literature reviews and scholarly commentary like essays or editorials.

[3] http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/3240

Descendent publication: A descendent publication is a paper that cites a reuse publication. If a dataset A is reused in paper B, and then paper C cites paper B, paper C is a descendent publication of dataset A.

Potential reuser: A potential reuser is an individual interested in reusing a dataset, but who has not yet engaged fully with the process. He or she may be developing a research question, choosing which (if any) dataset to work with, etc.

1.7 Contribution and significance of the study

Together, the three investigations that comprise this study represent a comprehensive exploration of how to identify, acknowledge, and anticipate data reuse in the social sciences. The first part addresses the question of how to reliably identify data reuse in the literature. A fundamental problem underlying the ability to formally recognize the contributions scientists make when they make their data available for reuse is the current lack of knowledge of how to identify instances of reuse; the first part of my dissertation addresses that gap by establishing how and why social scientists cite data. Building on that, the second part of my dissertation speaks to the problem of what formal recognition of data impact should look like. By proposing several methods of measuring reuse impact, some of which hew closely to traditional impact measures and others that go beyond those measure to account for alternative kinds of impact, I provide a proof-of-concept for quantifying reuse impact as well as concrete suggestions as to the best metrics. The final piece of the dissertation builds on these earlier sections but provides an alternative perspective. An implicit but enduring question when planning for the archiving and preservation of materials is whether the item is important to preserve or not. The final section of my dissertation examines the relationship between characteristics of the data, their producer(s), and their documentation to identify what (if any) of those characteristics are associated with high levels of reuse or high impact. Understanding

what leads data to be reused frequently or to be reused in important ways will help curators, funders, and individual researchers make decisions about what and how to curate.

I consider the primary contributions of this study to be the following:

1. Developing and testing a set of reuse impact measures that can identify important data sharing contributions.
2. Identifying factors that predict whether a dataset is likely to (1) be reused and (2) have high or low impact.
3. Creating a set of impact measures that present alternative interpretations of impact beyond measures based on citation counts.

The study speaks to the interests and needs of three major stakeholders: repositories, funders, and individual researchers. Better data impact measures have the capacity to identify scientists who have already shared important data as well as the repositories that have curated, preserved and made that data accessible. But the value of developing a better understanding of how to quantify the impact of data is not only in the ability to recognize key contributions retrospectively. One barrier to making datasets available for reuse is the lack of incentives for data producers to do so. Better metrics for data impact will help researchers identify and promote the value of their scholarship and may hasten the acceptance of data publication as a legitimate research output.

In addition to developing and testing metrics for data impact, this study also explores ways to identify data that are likely to be of high value. Repositories need to allocate their resources wisely, and to do so, they need to be able to make informed decisions about whether data will be valuable in the future. This study provides repositories with some

heuristics to guide their collection development and curation strategies, particularly through monitoring data producers' output. A key indicator of whether the datasets in my sample were likely to be reused or not was the number of publications put out by the data producers; repositories may be justified in recruiting data that are heavily published more aggressively than data that do not result in many publications.

By allowing repositories to enumerate the total impact of the data in their holdings, data reuse impact metrics may help repositories as they solicit funding, data, and users. The final group of stakeholders that the results of this study speak to are funding agencies. Funders wish to distribute their funds in a way that maximizes impact. A more nuanced understanding of how data contributes to scholarship will help funders direct money to those researchers and projects that are likely to make the most difference.

Finally, this study will also contribute to the field of scientometrics. Existing impact metrics focus primarily on publications and individual authors, and they consider almost exclusively only citations directly to a work. This approach is insufficient for data, given the ways that data citation patterns differ from citations to publications. The new metrics for data that I introduce and test here contribute to a more nuanced, broader understanding of scholarly impact; though these metrics are targeted directly to data, they may also be adaptable to other contexts. The availability and acceptability of a broader range of impact metrics undermines some of the "perverse incentives" described above that lead to a publish-or-perish mentality in academia by recognizing and highlighting different kinds of impact.

1.8 Overview of the dissertation

The next chapter provides an overview of the literature relating to this dissertation: specifically, previous work addressing incentives and barriers to data reuse, data citation practices across the sciences and social sciences, and how scholarly impact can be measured and anticipated. Chapter 3 follows from literature review to describe the overall methodology and approach for this dissertation, as well as describing the sample of datasets that form the basis of the study.

In Chapters 4 through 6, I present my findings in three parts. My primary research question is divided into three subquestions, each of which is addressed in a separate chapter. These three chapters follow a similar format: in each, I present (1) the specific methods for that chapter, with relevant literature; (2) that chapter's findings; and (3) a discussion of those findings. Chapter 7 provides a summary of all the dissertation's primary results along with a discussion that connects the three chapters to the primary research question. I conclude with an examination of the limitations of this study and directions for future research.

CHAPTER 2

Literature Review

In an ideal world, any data that had high value for sharing would be shared. Repositories and other interested parties would be able to accurately assess the value of the data and develop preservation and curation strategies accordingly. They could then provide the data to reusers, along with all the information reusers would need to effectively work with the data, and the scientist who created those data would get credit for the contribution they made through their sharing. Unfortunately, there are significant barriers that preclude the realization of this ideal world, and further, there are gaps in our knowledge that prevent the breaking down of these barriers. This literature review focuses on three areas related to data sharing at different points in the process. First, I address incentives and disincentives for sharing data. What prevents data from being shared? What changes could facilitate more effective sharing? Second, I turn to practices around data that are already shared, specifically how data reusers use (or fail to use) citations to acknowledge data producers. Finally, I explore the problem of identifying data that are likely to be reused.

2.1 Incentives and Disincentives for Sharing Data

Data sharing has a wide variety of benefits for scientists, including increased efficiency in the research cycle; new research capabilities, including the ability to mine shared data from repositories; effectiveness, especially the increased potential for wider scrutiny of

research results; and knowledge exchange and impact (Whyte & Pryor, 2011, p. 206). There are benefits for funders as well, particularly in the savings that can be realized if scientists are able to reuse data instead of re-collecting them. Yet despite widespread recognition of the potential advantages of sharing data, there is abundant evidence that scientists continue to withhold data, even in fields that are upheld as exemplars of openness.

Genomics, for example, has a long-standing data repository, and many journals have long required scientists to submit sequences to GenBank upon publication of results. A 2006 study showed, though, that between 3% and 15% of studies do not comply with the requirement (Noor, Zimmerman, & Teeter, 2006, p. 1113). A later study in a related field confirmed these results, finding that deposit rates of gene expression microarray datasets was quite low, less than 50% (Ochsner, Steffen, Stoeckert, & McKenna, 2008, p. 991). The picture in the social sciences is similar. A survey of Canadian researchers in 2008 found that less than half of their respondents said they had plans to archive their data (Perry, 2008). The subjects interpreted 'plans to archive' broadly, though: their plans included everything from destroying their data to keeping them in their office to depositing them to an archive. Only 19% of the respondents had identified an archive to which they could send their data. Among subjects who had received government funding, the picture was somewhat rosier, with around 65% reporting plans to archive their data (2008, p. 139). Though it is true that social science data are more often shared informally than through a repository (Pienta, Alter, & Lyle, 2010), it is clear that many subjects' 'plans to archive' outside of a repository could introduce barriers to sharing.

Some scientists, especially in the life sciences, fear being ‘scooped’ or losing the ability to continue to publish from the data once they are released to a broader audience. This fear is not entirely unfounded: around 35% of geneticists in one survey reported that they had been scooped after sharing information before publishing their own work (Louis, Jones, & Campbell, 2002). At the same time, though, larger proportions of respondents in the same study experienced positive outcomes from sharing their data, from collaborations and additional publications to new research and new grants. A 2003 report laid out several reasons for data withholding in the life sciences, highlighting in particular the role of the for-profit sector in such research. Increasing amounts of biological and chemical research occur within private organizations, which, for competitive reasons, may be reluctant to share data, regardless of the desires of the individual scientists carrying out the work (Committee on Responsibilities of Authorship in the Biological Sciences, 2003, p. 22).

Among the gene expression microarray scientists Ochsner et al. (2008) studied, scientists identified the primary barrier to depositing data as the effort required to prepare data. Since microarray data require a more complex metadata structure than gene sequence data to make them understandable and usable by others, scientists have to do more work to make them shareable and thus they are shared less often. Close to 80% of academic geneticists in a 2002 study rated “Effort Required to Actually Produce Materials or Information" as an important or very important reason to deny requests from their colleagues and other researchers (Campbell et al., 2002).

While the benefits of data sharing are widely known, and many funders, policy-makers, journals, and institutions along with some individual scientists push for greater participation in data sharing, many scientists still withhold their data. The reasons for withholding are myriad, as are the approaches put forth to counteract them. A crucial barrier is that sharing data adds to scientists' workloads. Scientists find data sharing effortful because of the relationship between data and the scientific process. Data are not isolated or static objects; instead, they are constantly interpreted and reinterpreted as a research project evolves (Hilgartner & Brandt-Rauf, 1994). Scientists use "common sense categories" to describe data at different stages of production, analysis, and dissemination (1994, p. 359). This complicates the question of how to make data shareable because it adds a new dimension: *when* are data shareable? Sharing data is not simply a problem of finding the appropriate moment or the appropriate product to pluck out of a research project and deposit to an archive somewhere.

In order to be shared, data must be separable from the process in which they were created; for some scientists, this happens naturally, but for others, doing so requires a radically different orientation toward their work and imposes a significant burden (Carlson & Anderson, 2007). Data, in order to be shared for reuse, "need to be rendered disseminative, that is to be rendered at the same time transportable in concise abstract forms and intelligible" (Carlson & Anderson, 2007, p. 641). Producing these disseminative forms often requires integration of contextual information with the data itself, especially through documentation describing details of the data collection, the

research design, etc. As a result, data that are seen as the most 'shareable' tend to be those which are easiest to share, rather than those which might have the highest value to other researchers (Cragin, Palmer, Carlson, & Witt, 2010, p. 4036). For some scientists, making data concise, abstract, and transportable happens in the course of ordinary work. When large project teams are involved, each member has a strong incentive to clearly and explicitly document the work that they do, the conditions under which it was done, and its relation to other parts of the project (Carlson & Anderson, 2007).

Big science fields, which have larger communities of researchers and broader funding bases, such as climate science, and those that rely heavily on secondary data use, such as epidemiology, have built on this strength and made significant efforts toward enabling data sharing. These efforts often focus on creating metadata for archiving and sharing data. Lawrence et al. describe their work in creating a metadata infrastructure for the NERC DataGrid, a data infrastructure that provides services to scientists funded by the UK's National Environment Research Council. They developed an "information middleware" comprising a taxonomy of eight classes of metadata, including metadata for archiving, for browsing, discipline-specific metadata, and ontology metadata. What they refer to as 'archiving metadata' is key to sharing; these metadata "provide a complete description of the data, constructed in a manner which can, where appropriate, be used by different communities to obtain the data in their own native formats" (2009, p. 1007).

However, other fields have found barriers to this kind of comprehensive metadata creation process: researchers can be resistant to spending additional time and effort to

create their own metadata, and researchers can also be unprepared to address issues such as intellectual property concerns, privacy, and concerns over data quality (Mathys & Kamel Boulos, 2011). In other fields, especially those like anthropology, in which researchers work primarily on their own, this kind of extensive self-documentation is neither the norm nor is it necessary to doing successful work (Carlson & Anderson, 2007). For scientists in such areas, incorporating the kind of activities necessary to enabling reuse "would imply a radical epistemological change" (Carlson & Anderson, 2007, p. 642). The documentation many scientists produce in the course of their work is geared toward providing themselves with the contextual information they need to fulfill their own research questions. As Shankar (2007) notes, the act of creating records is situated in the specific contexts in which scientists work; their records attest to an activity that is only meaningful within that context.

Scientists manage data for their own use and to suit their own purposes; the actions they take while producing data are those that best fit their workflows. Where these actions produce shareable data, it is sometimes seen more as a positive side effect rather than a goal in and of itself (Fear, 2011). A primary goal of scientists throughout their work is to produce data and results that generate a publication. Borgman et al. (2007) found an intimate relationship between data practices and publication practices. Their subjects had the publications they would eventually produce foremost in their minds during the data collection process; the scientists they studied "design a field experiment with a particular story in mind, and […] the story determines roughly the amount of data needed" (2007, p.

26). The particular configuration of data necessary to produce the desired publications is, in a sense, baked in.

When the work to produce data is especially onerous, the additional work required to turn that data into something that can be shared may be even less palatable (Birnholtz & Bietz, 2003; Borgman et al., 2007). This results in a conflict between the practices that best suit the everyday work of science and those that promote the creation of shareable, reusable data (Akmon, Zimmerman, Daniels, & Hedstrom, 2011, p. 331). Information that is extraneous within the context of a particular project may be crucial to reuse of those data for another purpose. 'Everyday work,' though, is changeable, even within the practices of an individual; as individual scientists' research interests and needs shifted over time, so did their documentation practices, which makes it difficult for others to understand and access the data for other purposes.

Sharing data requires scientists to create a new kind of communicative product. No longer are data simply a means to a paper; data sharing makes data an end in and of itself. Asking scientists to share their raw data constitutes a different kind of scientific communication, one which scientists may not be entirely prepared or easily convinced to engage in. Scientists may be reluctant to engage in additional work, unless a very clear case can be made for how that additional work will benefit them (Pritchard, 2005). For this reason, simply having tools such as large, open-access databases available to facilitate sharing is no guarantee that sharing will take place (Akmon, 2011).

Numerous approaches have been taken to encourage scientists to share their data, from requirements by funders, journals and other authorities, to appeals to scientists' sense of duty and responsibility toward openness in science. Mandates from journals appear to be an especially effective approach. Vines et al. (2013) found that a strong mandate—one that requires the inclusion of a data availability statement in a manuscript—increases the likelihood of data sharing by a large margin. In political science, journals are increasingly adopting data sharing mandates, particularly higher impact journals (Gherghina & Katsanidou, 2013).

The authors of a 2003 report on sharing in the life sciences identify data sharing as an "obligation" of data producers; not only should scientists release data that is fundamental to a paper, they should also "provide them in a form on which other scientists can build with further research" (Committee on Responsibilities of Authorship in the Biological Sciences, 2003, p. 4). While noting some structural changes that would facilitate scientists sharing data (for example, funders providing appropriate funding to enable data sharing, better acknowledgement by data users, and reviewers specifically requesting material that may not be initially included with a manuscript), the authors place the primary responsibility and burden on scientists themselves to make their data not only available but useful to others. Data sharing mandates from the NIH (National Institutes of Health, 2003), the NSF, and several prominent journals similarly place the burden of data sharing on scientists who receive funding from these agencies or who hope to publish in these venues.

Such mandates represent the 'sticks' in the effort to foster widespread data sharing, and suggestions of appropriate 'carrots' have proliferated as well. One approach is to safeguard the value data producers get from their own data by allowing for embargo periods after data are deposited to a repository. Another approach focuses on findings ways for data sharing to produce even more value to data producers. Developing norms for data citation, for example, to acknowledge the scholarly contributions of data creators could incentivize scientists to engage with the additional work of data sharing (Lord et al., 2005). Kaye et al.'s (2009) editorial highlights the importance of researchers gaining recognition for the work they do, and they note that punitive mechanisms for dealing with scientists who decline to share their data are "inappropriate and cumbersome." Instead, positive rewards for sharing may be more effective in increasing scientists' compliance with data sharing initiatives. Data sharing constitutes the creation of crucial resources for other researchers, and that contribution should be recognized as such.

The idea that data could be an important resource to others is an especially attractive incentive for sharing. A majority of respondents (65%) in a 2008 study said that they were more likely to deposit data if they thought those data would benefit other people (Hedstrom & Niu, 2008). But many scientists feel that their data does not or would not have value to others. Researchers at a UK university reported that they expect nearly half of all their data to have a useful life of fewer than 10 years, and just 27% might have value over the long term (Baudoin, 2008). There is a chicken-and-egg problem here, though. While this estimation may be grounded in concrete knowledge about obsolescence of data or the instruments used to collect it, some scientists may believe

their data are not valuable to others simply because they are not aware of what others could do with them. If data were shared more widely, more scientists might recognize the possibilities of unexpected uses for their data, and then they might be more likely to share.

Exposure to the benefits of data sharing does seem to influence whether someone shares her data. In genomics, the scientists who are most likely to share their data are those who have published in an open access journal, have previously shared gene expression data or have previously reused gene expression data (Piwowar, 2011, p. 7). Scientists who have reused data have directly benefitted from data sharing; those who have shared previously may also have indirectly benefitted from the knowledge that others have used (or have the potential to use) their data. A higher level of experience in research also correlates with data sharing. Data produced by researchers who are more senior, have published more, or been cited more are more likely to be shared than data produced by those who are less experienced (Piwowar & Chapman, 2010). Scientists with more experience in research may be better acquainted with the benefits of data sharing: they may have shared data informally or requested data from others, or they may be aware of others' efforts to share and reuse data.

If direct or indirect experience with data sharing and reuse bolsters willingness to share, better mechanisms for identifying and communicating successful instances of reuse and sharing might encourage scientists to share data. Evidence of the benefit data give to other people might lie in the amount of citations those data garner; if people are using

data from a repository, it suggests that the availability of those data saved others the trouble of collecting the data again (a factor that 76% of Hedstrom and Niu's respondents responded positively to). For data sharing to succeed, the incentive structure around data sharing must foster a culture of sharing. The broader the awareness of the benefits—direct and indirect, intrinsic and extrinsic—the more likely scientists are to share their data. Recognizing this, Piwowar et al.' s foremost recommendation for fostering a data sharing culture is for academic centers to "measure, recognize and reward" researchers' efforts to share their data (2008, p. 1315), particularly during hiring and tenure review processes. One suggestion they had was to let data sharing contributions supplement a scientist's impact factor; if a scientist shared data related to a paper that contributes to their impact factor, the influence of that paper should be given more weight.

Sharing research data is, in fact, associated with both an increase in the citation rate for the paper describing the data, independent of the impact factor of the journal, the date of publication, or whether the author is in the US or not (Piwowar et al., 2007, p. 2) as well as with a higher number of publications overall from a dataset—both by the original producers of the data producers and by data reusers (Pienta et al., 2010). These findings indicate a second benefit of measuring data reuse. In addition to exposing scientists to the ways in which they can help others by sharing their data, measuring data reuse enables scientists to highlight and quantify their own contributions. Scientists seek credit for their work not just to reap the rewards of their accomplishments, but also to build their credibility as scientists (Latour & Woolgar, 1986). In this way, acknowledgement for data sharing bridges intrinsic and extrinsic incentives. Citations to data, as a measure of

how much use a shared dataset gets, can be a powerful tool for communicating to a hiring or tenure committee about the value of a scientist's work. But citations can also serve as evidence or feedback to the data producers that their data were not only used but also benefited other scientists in their work.

Intrinsic and extrinsic rewards can work hand-in-hand in science. Particularly, where informal collegial recognition of contributions is lacking, scientists are more heavily influenced by external rewards: formal rewards like citation counts are strong drivers of research productivity where there is an undeveloped or non-existent culture of informal rewards (Reskin, 1977). Currently, in many areas of science, there are no structures to reward data sharing either formally or informally. Philip Bourne, Editor-in-Chief of *PLoS Computational Biology* and co-director of the Protein Data Bank (PDB), uses the following anecdote to illustrate the mismatch in incentives for sharing data:

> The structure of human deoxyhemoglobin is one of the most downloaded structures in the PDB—in one year, it has been downloaded more times than the original paper has ever been cited thus far. Yet from the authors' perspective, the Nobel Prize does not come from constructing the PDB database entry, but from an eloquent description of the relationship between structure and function that was presented most completely in the literature. A tenure committee does not award tenure based on the number of deposits a faculty member has made to a biological database, but rather the number of papers they have published in leading journals. (Bourne, 2005, p. 180)

McDade et al. (2011) forcefully argue that the current metrics for measuring researcher productivity in systems biology are insufficient. Modern systems biologists contribute to databases, maintain web sites that host their research results, develop laboratory or software tools, pioneer interdisciplinary projects, and engage in many other non-traditional forms of scholarship. But these contributions are ignored by the most broadly

accepted measures of research output to the detriment of innovative, cutting-edge scientists (2011, p. 619). They also note that metrics, even new metrics, based solely on the quantity of citations a work of scholarship attracts unfairly favors scientists who work in popular areas, echoing Bowker's (2000) comments on the outsized influence of 'charismatic megafauna' on biological databases.

The work of data management is rewarded indirectly through journal publications, or more precisely, data management work that contributes to a publication is rewarded through the successful completion of a journal paper. Despite the citation advantage that arises from data sharing, the work to publish data in and of itself has no direct rewards; feedback from users can be heartening to a scientist, but without formalization, the independent impact of a shared dataset cannot be effectively communicated to tenure or hiring committees. If publishing data were rewarded, however, this could be a powerful force towards widespread data sharing. More specifically, if data sharing were seen as a legitimate scholarly contribution, it could encourage scientists to publish their data, thus exposing them directly to the rewards of data sharing and further encouraging their future sharing behavior (Nelson, 2009).

2.2 Identifying Data Reuse

There is an important mechanical issue underlying the problem of incentivizing data sharing: in order to reward a scientist for the contribution she makes in sharing her data, other scientists' reuse of that data must be visible. In other words, crediting data reuse necessitates an unambiguous linkage between a data producer, a dataset, and the context of reuse.

A problem arises when researchers fail to show clear linkage between results documented in a paper and the data used to produce those results (King, 2011). While large datasets are reliably archived and thus are generally findable, small datasets are often not locatable, either because they have been lost or the producer chooses not to release them (or to share only with a select few individuals). Altman and King (2007) succinctly describe the problems that the lack of standardized data citation practices causes throughout the sciences:

> The data cited may no longer exist, may not be available publicly, or may have never been held by anyone but the investigator. Data listed as available from the author are unlikely to be available for long and will not be available after the author retires or dies. Sometimes URLs are given, but they often do not persist. In recent years, a major archive renumbered all its acquisitions, rendering all citations to data it held invalid; identical data was distributed in different archives with different identifiers; data sets have been expanded or corrected and the old data, on which prior literature is based, was destroyed or renumbered and so is inaccessible; and modified versions of data are routinely distributed under the same name, without any standard for versioning. (n.p.)

If data are not linked to the papers in which they are used, it is impossible to identify and locate the data used in the analysis, and that analysis is not replicable. Second, when data are not visible to the public, it presents a barrier to giving the producer or creator of the data credit for their production. The problems of data access and assigning credit for data are linked through the issue of incentives: where there is no mechanism for recognizing contributions made through sharing data, there is little incentive for scientists to do so.

Researchers have argued that the idea of a journal publication as the standard unit of communication is no longer effective in today's scholarly communications landscape

(Van de Sompel, Payette, Erickson, Lagoze, & Warner, 2004). They cite alternative means of publication and communication such as simulations, software, and data that should be acknowledged as publications in their own right. Callaghan et al. (2012) distinguish between 'publishing' (lower-case p) and 'Publishing' data. Small-p publishing data is easy: anyone can put their data on a website. But big-P Publishing is a challenge. This matters because Publishing carries with it a set of guarantees about quality and longevity that ad-hoc, small-p publishing does not. Formal publishing also carries a set of incentives that are lacking in informal publishing: formally published objects can be cited and tracked, and credit and recognition can be doled out accordingly. While this is not true for all disciplines, expanding what signifies as publication and developing means of tracking these alternatives are key for any system of incentives to grow.

With this in mind, Chavan and Penev (2011) note that one method for making data citable and discoverable is to support publication of 'data papers'. The data paper is a publication whose primary purpose is to describe data and links directly to a dataset. This idea has been taken up by the National Environment Research Council (NERC) data centers. Callaghan et al. describe the work the NERC has done to formalize data publishing via a data paper, "which acts as a proxy for the dataset and describes the technical form and scientific content of the dataset, and acts as a guide to its use for other researchers" (2012, p. 112). Data papers could appear in standard journals, or in data-specific journals[4] (UC Curation Center and Publishing Services, California Digital Library, 2011). This publication model brings the benefits of formal publishing to data,

[4] For a current list of data journals, see http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList

and crucially, makes the data into an object easily recognizable as a citable entity. If the data look like a paper, they are (in theory) simple to cite.

While the approaches described above capitalize on new technical capabilities (especially the affordances of electronic publishing) by introducing entirely new modes of publication, these authors propose significant changes to the existing system of publishing scientific results. There is, however, a long-accepted method of linking a publication to external information via citation. A simpler solution to the problem of linking data to publications and ensuring both data discoverability and acknowledgement of the data producer is to capitalize on existing norms for acknowledging scholarly debts and identifying sources through standardized citations. A 2012 workshop run by the CODATA-ICSTI Task Group on Data Citation Standards and Practices explored this issue in some depth and across a range of disciplines (Uhlir, Board on Research Data and Information, Policy and Global Affairs, & National Research Council, 2012). That group found that citation practices are uneven across fields and that there are barriers, both technical and social, to the development of common, standardized practices in the scholarly community. They emphasized, however, the importance of citations for attributing data.

Citations already serve multiple roles in papers. Peritz (1983a) presents a classification of citation roles for the social sciences and other closely-related fields. His schema comprises eight categories. Citations function (1) to 'set the stage' for the citing study; (2) as background information; (3) to support or describe the citing study's methods of

design or analysis; (4) to provide a comparison to existing work; (5) to support the formulation of new hypotheses or conjectures, suggestions of future research, etc.; (6) to point to the study's data sources; (7) to trace the history of a subject or to highlight the work of pioneers in an area; or (8) as off-hand references to related work. His analysis of five journals, one each from five fields (sociology, epidemiology, library and information science, education, and demography) found that the most numerous kinds of citations were setting the stage, followed by methodology. Peritz noted that few of the articles cited documentary sources in any of the five areas, except in library science. Less than 6% of the total citations in each journal indicated raw data sources, which may relate to the prevalence or rarity of secondary analysis and reanalysis in the sample of journals Peritz examined.

Formal data citations would fit into Peritz's category (6), documentary citations. Data citation has the potential to transform work that many scientists see now as a purely internal "housekeeping chore" (Walton, 2010) by providing external recognition of that work. In his editorial, Walton stressed the importance of data citation, identifies progress toward enabling and supporting data citation, and announced that his journal (Antarctic Science) would begin encouraging authors to include data citations in their work. At the same time, Walton noted that much remains uncertain about how such a policy will be implemented: standard citation formats, or standards for who exactly to cite (the producer, the provider or both), or even standards for where and how the citation should appear (in the text? in the references?) need to be established before data citation truly becomes the "new norm."

There have been numerous efforts to define data citation standards, not all of which agree on the elements to be included. In the environmental sciences, Lawrence and colleagues identified problems with the piecemeal approach taken to developing standards for data citation—for example, they criticized the American Geophysical Union's insistence that the physical location of a data center be reported in a data citation (Lawrence, Jones, Matthews, Pepler, & Callaghan, 2011, p. 25)—and recommended syntax of their own. The International Polar Year's data policy explicitly defines a format for citing data which acknowledges both the data producer as well as the repository from which the data were acquired (Parsons, Duerr, & Minster, 2010, p. 298). Altman and King (2007) take a broader approach and propose a standard format for citing quantitative data. Their standard is purposefully minimal, to make it easier to use: the required elements are simply the name of the data producer(s), the date the data were published, and the title of the data, along with a unique global identifier, such as a DOI, a Unique Numeric Fingerprint (a short numeric code uniquely tied to the content of the dataset, which would allow for version control), and finally, a bridge service component, such as a URL, to enable easy access to the data.

A common element among these proliferating standards is the need for a unique identifier attached to the data. DataCite is an international consortium to support publication and citation of data. They provide persistent identifiers in the form of DOIs to enable standardized citation of data and the ability to track data citation over time (Wilkinson, 2011). An alternative approach is the use of numeric ID numbers. When citing genomic

and proteomic data, for example, authors rely on GenBank accession numbers (Brown, 2003, p. 930), which in some cases, have been retroactively inserted into existing papers.

Two elements of prior research suggest that the social sciences in particular would be a fertile area into which to introduce formal data citations. First, Peritz, in a different study, found that the most commonly cited works in sociology were methodological papers (Peritz, 1983b, p. 217). Henige suggests that authors in the social sciences are relatively more thorough with their citations, at least compared to authors in the physical and life sciences, particularly with respect to their inclusion of page numbers for direct quotes. Where this practice is "below the radar" in the sciences, the provision of precise citations is of some interest in the social sciences (Henige, 2006, p. 106). Together, these findings suggest that the social sciences have existing norms for both comparatively thorough citations and for clear acknowledgement of usage of materials and methods developed by others. Adoption of a formal data citation standard, then, would be a natural extension of these existing norms.

In fact, major social science data repositories do have well-defined citation standards and guidelines to help authors use them.[5] Despite this guidance, though, data citation has not fully been incorporated into practice. Sieber and Trumbo (1995) looked at citation patterns among a sample of 198 papers that used data from the General Social Survey. Their work paints a very negative picture of the state of data citation in the social

[5] For example, see ICPSR's Data Citation page (http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/citations.jsp) or the Roper Center's guidelines (www.ropercenter.uconn.edu/data_access/data/how_to_cite.html)

sciences. In their sample of secondary papers, only a fraction provided a formal citation to either the data (19%) or the data producers (13%) (1995, p. 14). Further, their sample covered a period of 12 years (1976-1988), and there was no evidence of improvement in data citation practices over that span of time (1995, p. 14). Perhaps most troubling, their survey of authors showed that 81% would not specifically acknowledge the producer of the data they used (p. 17). A minority (11%) reported that no acknowledgement of the producer would be necessary when a repository provided the data on the producer's behalf, or when the data were in the public domain (1995, p. 17).

In a similar, later study, Mooney examined 49 secondary papers from the ICPSR data-related literature bibliography published between 2000 and 2009 to determine if and how their authors cited the data they used. Like Sieber and Trumbo, she found that the majority of the papers—about 60%—failed to include any kind of citation to data (Mooney, 2011, p. 103). Of the papers that did provide a citation of some kind, though, most included a formal citation in their reference list. Though the proportion of papers citing data here is paltry, there is some evidence that even this low rate is higher than found in other fields. Parsons et al. (2010) include in their paper a “quick” analysis of papers using data from the Snow and Ice Data Center that shows that less than 10 percent of the papers identified included a formal citation to the data used.

One possible reason for the lack of data citation is that authors may primarily cite papers written by the producers of the data they use. Brown (2003), studying genomic and proteomic data reuse, made an attempt at conventional citation analysis of data reuse in

this field which failed because references to genomic and proteomic databases were inconsistently placed within manuscripts, making it difficult to carry out automated tracking of citations. She also found that it was common for authors to cite the paper in which data first appeared and in which the accession number for the data is assigned rather than the accession number itself. Mooney (2011) found that roughly half her sample cited data producers' papers in addition to or instead of citing data directly.

The lack of data citations is problematic because, as Mooney points out, failing to acknowledge a scholarly debt is poor practice and borderline unethical. But there are practical implications as well, specifically in the inhibition of metrics for data impact. Because the fundamental unit for most impact measures is the citation, inconsistent data citation practices pose a significant barrier to accounting for data reuse impact.

2.3 Measuring Data Reuse Impact

The suggestion that data sharing should be rewarded raises the question of what metrics should be used, or more precisely, what exactly should be counted when it comes to data impact. Traditional measures of impact—which address the scholarly impact of journals, articles, and researchers—are based on citation counts. Setting aside, momentarily, the issue of how well or consistently scientists cite data, a further question remains: are citations the best fundamental indicator of a dataset's reuse impact?

To begin with, I review traditional measures of research impact. The earliest work on research impact aimed not to assess individual pieces of scholarship, but instead to develop an objective system to rank journals in terms of their importance to their fields.

The most commonly used measure of impact is the impact factor, which generally applies to journals but has also been extended to authors (Garfield, 2006). The journal impact factor measures usage (how often other papers cite articles that appear in a given journal) balanced against the amount of substantive work the journal publishes. Typically the calculation is limited to publications and citations within the previous two years. This ensures that a particularly well-received paper or set of papers does not artificially inflate the journal's impact over time. The formula works very well within fields but somewhat less well when comparing journals across fields: different fields have different half-lives (the number of retrospective years needed to be included to reach 50% of cited references), so the two-year interval can cause one field's journals to have a depressed or inflated impact factor related to another field's journals.

Journal impact factor does not, however, correlate well with citations to individual articles within journals (Seglen, 1997); it is not a good metric for measuring article-level impact, despite its occasional use as a measure of research productivity. To measure individual research impact more effectively, Hirsch (2005) proposed the h-index, which assigns an individual an index *h* where h represents total number of papers with h or more citations, and the rest have fewer than h citations. Advantages of this measure are that it provides a rough estimate of the total number of citations a researcher's papers gather, and that it can only increase or remain the same over time, which preserves a sense of a researcher's impact over time. Further, the h-index cannot be skewed by especially high- or low-performing papers. That said, the lack of sensitivity to high-performing papers can be considered a drawback; papers that have an outsize impact on a

field should arguably have an outsize impact on the measure of the author's productivity. For that reason, Egghe (2006) introduced the g-index, which adapts the h-index to account for the accumulation of new citations to an author's highest-performing papers.

Moving away from the researcher him- or herself, one can consider how to measure the impact of a singular piece of scholarship. Several possible alternative metrics have been proposed which can potentially be applied to individual datasets. Antelman (2004) uses *citedness*, the total count of citations to a given paper. Citedness is not an exact analog of impact, but "is itself viewed by scholars as an objective" (2004, p. 374), which makes it a potentially important measure. Other work on individual impact goes beyond citation counts. Bollen et al. (2008) proposes a shift to usage-based impact, where usage is defined more broadly than citation. Their MESUR project draws on a range of usage events corresponding to different stages of the research process: "throughout the scholarly process articles are discovered, downloaded, e-mailed to peers, read and saved for later consultation" (Bollen et al., 2008, p. 231). Similarly, the COUNTER[6] project facilitates tracking of usage metrics of online materials; by recording the various ways in which users interact with electronic resources, both librarians and publishers can reliably report usage of their collections or materials and in turn derive additional metrics such as cost-per-use or explore usage patterns. Records of these actions are not subject to the same delay that citation data is, and these actions capture usage of articles that may not appear in a formal citation.

[6] http://www.projectcounter.org/about.html

These measures of impact easily lend themselves to data. Conceptually, the research process is similar when building on ideas from someone else's publication or on their data. In both cases, a scholar produces new work from old, and in a perfect world, he or she would acknowledge that prior work through citation. Using the ICPSR Bibliography of Data-Related Literature, it is possible to assess the citedness of each dataset in the repository, and more specifically the number of citations that represent reuse (i.e., citations from authors other than the data producers). And while ICPSR does not track all of the usage events the MESUR researchers had access to in their study, the repository does maintain download statistics. Download statistics may capture usage that citation counts do not, for example student work which is never published.

Arguably, there is a huge range of ways in which research can make an impact, many of which are never reflected in a work's citation count or even the number of usage events it produces. These are what the Australian Research Quality Framework refers to as "extra-academic impacts," or the "wider economic, social, environmental, and cultural benefits of research" (Donovan, 2008, p. 47). Understanding such impacts can be a boon to researchers (who need to make the case to funders and to their institutions that their work is worthwhile) as well as to funders, institutions and repositories, all of which could benefit from a more thorough understanding of how research makes a difference. Problematically, though, this kind of impact is much more difficult to count and measure.

Lavis et al. (2003) proposed a method of identifying and quantifying such broader impacts. They identify three kinds of research impact measures: producer-push (effects of

the researcher's active actions, such as number of papers produced, the status of the journals in which those papers are placed); user-pull (results of users' active efforts to identify and use research, such as citations to papers); and exchange (joint efforts of users and producers to apply research and to shape research agendas, for example through an NSF workshop). Traditional impact measures like impact factor and h-index draw on producer-push measures and a limited subset of user-pull measures. The value of research, they suggest, may be better understood by emphasizing a broader set of user-pull and exchange measures, i.e. "[m]oving beyond *whether* research was used to examine *how* it was used" (2003, p. 165). Lavis et al. were especially interested in translational work produced from health research, so their measures focus on the penetration of research into decision-making spheres, including things like indications of decision-maker awareness of a specific piece of research and the number of meetings decision-makers request with the group of researchers.

Central to Lavis et al.'s ideas about impact is the notion of knowledge transfer. Research has an impact when the ideas, findings, and knowledge embedded in the research travel beyond its initial publication. Citedness is one aspect of this. The citedness of data, as described above, indicates how often the data are reused in subsequent publications. Reuse in and of itself, though, may not mean much if it does not result in some kind of knowledge transfer. Considering, instead, knowledge transfer stemming from data reuse has the potential to provide a more meaningful and nuanced conception of data impact.

In a recent report, Costas et al. (2013) provide a framework for thinking about data metrics. They identify two types of metrics: size dependent and size independent indicators. Size dependent indicators, also called total performance indicators, are metrics like citedness or the h-index, wherein impact is quantified through the raw performance of the unit of analysis. Size independent indicators are based on an average performance, rather than raw numbers. An example of this type of metric for a dataset would be the average number of citations its associated data publication receives, for example. While their report does not suggest any specific metrics for data, their description of the conceptual landscape for data metrics provides a strong foundation for future development of metrics.

The broad landscape they describe also highlights an additional issue. Single indicators can be misleading; more powerful results can be obtained when several different impact metrics are contrasted with one another (Abramo, D'Angelo, & Viel, 2013; Brembs & Munafò, 2013; Saleem, 2011). Further, authors can game metrics (Falagas & Alexiou, 2008), for example by publishing results across multiple papers to increase their citation count. Data metrics could be subject to the same problems, particularly if a single indicator becomes especially prominent. The best way to build on the strengths of impact metrics for data while guarding against their weaknesses and (to the extent possible) the unintended consequences that may arise from their use seems to be to triangulate impact as much as possible. Using multiple indicators derived from different conceptualizations of impact could give a more complete and more reliable picture of data impact than any single measure would on its own.

2.4 Predicting impact

Metrics are constructed to reward past achievement but are often used as an indicator of future potential, particularly in hiring and promotion decisions. Though there is no study documenting widespread use of impact metrics in hiring and tenure committees, counting metrics like h-index correlate strongly with hiring and tenure decisions in numerous fields (Ellison, 2010; Kelly & Jennions, 2006; Svider et al., 2013). A 2010 survey reported in *Nature* found that 70% of surveyed researchers believed metrics like the h-index are used in tenure and hiring decisions. This belief was contradicted by administrators, who said they put more emphasis on recommendation letters, but the article notes that metrics can appear indirectly in materials like recommendations, even if they are not explicitly requested (Abbott et al., 2010).

As is the case in evaluating researcher productivity, there are numerous stakeholders in data sharing whose concerns with data impact are both retrospective and prospective. Costas et al. (2013) interviewed funders, scientists, staff from libraries and data centers, individuals working with research infrastructure, and publishers to determine their perceptions on data metrics and found that they saw uses for data metrics to fostering sharing, to identify usage, to link data and publications, and to evaluate proposals, among other applications. These stakeholders may be able to use data metrics in ways very similar to how publication metrics are used now: not only to highlight previous accomplishments, but as a way to identify datasets (or scientists, or institutions) that show promise of future impact as well.

Though this kind of usage of impact metrics may be widespread already, there has been only limited work exploring whether impact metrics actually have predictive power. The h-index has been the most thoroughly examined in this context, though some attention has been paid to other factors, such as qualities of scientists' co-authorship networks (McCarty, Jawitz, Hopkins, & Goldman, 2012) and "research buzz" as measured by tweets (Eysenbach, 2011).

Hirsch, the creator of the h-index, found that h-index is a strong predictor of future success in research (2007). In a later study, Acuna et al. (2012) found that a scientist's h-index was predictive of her future h-index, especially in the short term, but was effective in the long-term only when combined with information like her total number of publications, the length of her career, and the number of papers placed in top journals.

At the article level, Didegath and Thelwall (2013) found that the number of citations an article received was primarily related to the impact factor of the publishing journal and the citation impact of the paper's references. Papers published in high-ranking journals and which cited highly-cited articles were in turn cited more often than papers in lower-rank journals or which referenced less often cited articles.

These studies offer some limited evidence that metrics – related directly or indirectly to the object of interest – can predict future impact. But there has been no such study for data impact (related, in part, to the lack of accepted metrics for data). It is unclear whether data should follow the same patterns as other scholarly works do: highly cited

papers tend to continue to be cited; highly ranked researchers tend to continue to be highly ranked. If the value in data is finite, though, this pattern should not hold true; at some point, the data will have been put to every use they possibly can be. But the value in data may not be finite, particularly if they can be recombined with other data to answer new questions. At the same time, other factors may be more important to the reuse of data than their current level of reuse.

2.5 Conclusion

While the value of sharing data is broadly recognized, that value cannot be unlocked if it is difficult or unrewarding for scientists to share their data. Currently, there is a lack of incentives for sharing data, and in particular, a lack of standardized metrics for identifying sharing and reuse impact. Mechanisms for highlighting the distinct contributions scientists make when they share their data are underdeveloped. But underlying that problem is the lack of standards for data citation. Scientists' contributions in sharing their data are not visible if they are not credited clearly when those data are reused. If citation could be standardized, there are possibilities for developing data metrics which could help change tip the balance for scientists in favor of sharing their data. To accomplish this goal, though, it is important to create a suite of complementary metrics that together encapsulate the many ways in which data can be valuable. Such metrics would be useful not only in fostering increased data sharing, but would also inform decision-making around funding and curation. Metrics not only identify prior contributions, but they can indicate future potential as well.

In the next chapter, I outline the design of a study to address these issues: What is the current state of data citation practice in the social sciences, and can we make it better? How can different conceptualizations of data reuse impact be quantified? Are there qualities of data or their usage that predict their future impact? Together these three investigations comprise a comprehensive exploration of how to identify and acknowledge data reuse in the social sciences. The first part addressed the question of how to reliably identify data reuse in the literature. A fundamental problem underlying the ability to formally recognize the contributions scientists make when they make their data available for reuse is the current lack of knowledge of how to identify instances of reuse; the first part of my dissertation addressed that gap by establishing how and why social scientists cite data. Building on that, the second part of my dissertation speaks to the problem of what formal recognition of data impact should look like. By proposing several methods of measuring reuse impact, some of which hew closely to traditional impact measures and others that go beyond those measures to account for alternative kinds of impact, I provided a proof-of-concept for quantifying reuse impact as well as concrete suggestions as to the best metrics. The final piece of the dissertation builds on these earlier sections but provides an alternative perspective. An implicit but enduring question when planning for the archiving and preservation of materials is whether the item is important to preserve or not. The final section of my dissertation examines the relationship between characteristics of the data, their producer(s) and their documentation to identify which of those characteristics are associated with high levels of reuse or high impact. Understanding what leads data to be reused frequently or to be reused in important ways

will help curators, funders and individual researchers make decisions about what and how to curate.

CHAPTER 3

Methods

In this dissertation, I examined data citations in the social sciences, measured the scholarly impact of data reuse as well as explored factors that are associated with whether a dataset is reused. The guiding question for this dissertation is:

> *What is the scholarly impact of data reuse? How can stakeholders anticipate the impact the data they fund, create, or curate will have?*

To address this large question, I broke it down into three related explorations. First, in order to quantify the scholarly impact of data reuse, reuse must be identifiable. I looked specifically at identifying reuse through data citation patterns. Though this issue has been examined previously in the social sciences, I present a larger study of citation patterns here as well as one that takes a more nuanced view of how social scientists use citations to acknowledge others' prior work on which they are building. Second, I developed a suite of impact metrics for data, which I applied to a sample of ICPSR datasets. By testing these metrics on a varied group of social science datasets, I was able to demonstrate their use as well as shed light on how these datasets can be high impact in different ways. Finally, I turned to the problem of anticipating reuse and impact and explored what factors correlate with reuse and with high scores on the impact measures I developed.

In summary, the specific research questions are:

- How and why do social scientists cite data?
- Which datasets held by ICPSR are high-impact according to different measures of reuse impact?
- What factors contribute to whether and to what extent a dataset will be reused?

Using a combination of bibliometric and statistical analysis, I (1) identified instances of data reuse in a corpus of social science literature; (2) determined the impact of the reuse of a set of social science datasets by applying traditional metrics of scholarly impact; and (3) identified factors that are associated with high levels of reuse and with high impact. These three explorations address the individual research questions, and altogether, they contribute to answering my guiding question.

3.1 Study overview

I designed my dissertation as a three-part study. In the first part, I examined the set of peer-reviewed or published papers related to a sample of social science datasets. Studying this corpus of papers allowed me to accomplish two primary goals: first, to identify instances of reuse of these studies, and second, to identify citation patterns in the social science literature. After identifying a subset of secondary publications (articles that reuse data) from the full set of data-related publications, I used that information to develop and test metrics of data reuse, which was the goal of the second part of my study. These metrics then served as the outcome variables in models I produced in the final section of my dissertation, which identified factors that are associated with data reuse and reuse impact. The following sections describe the parts of my study and how they fit together in more detail, along with key findings.

3.1.1 Identifying data reuse in a corpus of social science literature (Chapter 4)

A key precursor to quantifying and predicting the impact of data reuse is identifying instances of reuse. This part of the study focused on the question: how and why do social scientists cite data? There were two areas of interest in this part of the study. First, how do social scientists cite data? ICPSR provides boilerplate citation text, but do social scientists use it? Do they cite data producers' papers in addition to or instead of citing data? Where do data citations appear: in article text, in footnotes, in references, or some combination? Second, why do social scientists cite data? Do they cite only data that they use for a paper, or do they cite data for other reasons, such as highlighting prior work in an area?

The main data source for this chapter was ICPSR's Bibliography of Data-Related Literature. This database is a manually compiled list of articles, conference proceedings, theses, reports and other documents related to ICPSR's data. Focusing on the subset of data-related literature related to data in my sample (N = 2,323), I identified primary and secondary publications and manually coded each publication for how and why the authors cited data. An important finding from this process was that secondary users of data typically cited only papers written by the data producer in which they describe the data, rather than citing data directly or acknowledging the data provider. Where they did cite the data provider, it was generally in addition to citing the data producer. While on one hand, lack of consistent data citation is problematic, the consistency with which data reusers cite papers relating to the data suggests that reusers prioritize acknowledging the scholarly work of creating data over the work done to share them. I also found some

evidence that clear, prominent guidance on citing the data producer has a significant effect on data citations: when prompted to cite data directly, data reusers do so.

3.1.2 Quantifying the impact of data reuse (Chapter 5)

This section proposed and demonstrated the usage of four different methods for measuring data reuse impact. The primary question addressed in this section is: do different methods of measuring reuse impact show different datasets to be high impact? The four measures of impact present a range of interpretations of scholarly impact. Two measures, citation count and downloaders, follow traditional models of impact, where direct usage is the primary evidence of impact. The other two measures incorporate other information about usage: how high-impact the publications produced from a dataset are as well as how varied the fields that use the data are.

This section used the ICPSR Bibliography of Data-Related Literature and added additional citation data to it. In the previous section, I identified primary and secondary publications related to the datasets in my sample. I used this information to produce citation counts for each dataset in my sample. ICPSR also keeps download counts for each study, and I used that information to compute the second impact metric: downloaders. Using the primary and secondary publications from the Bibliography as seeds, I gathered bibliographic information from Scopus about the citation networks for each dataset. These networks provided the basis for computing the additional metrics of data reuse impact. Scopus also provided journal categorizations based on subject classifications, which I used to identify the field in which each citation occurred.

In this section, I computed each metric for each of the datasets in my sample, then used a statistical method to compare the rankings produced by each metric. I found that the four different metrics produced significantly different rankings of the datasets. This finding indicates that conceptualizing reuse impact different has a substantive difference on how datasets score: the most frequently used datasets were not always the ones with the most highly cited reuse publications or the ones used in the most diverse fields. No single measure was the best at identifying high-impact datasets.

3.1.3 Anticipating data impact (Chapter 6)

The final section examined whether there were factors that influenced whether a dataset was likely to be reused. Are there factors that are significantly related to the amount of reuse a dataset receives? Do the same factors relate to a dataset's reuse impact?

In Chapter 6, I used regression models to explore these relationships. I modeled reuse first as a binary outcome (were data reused or not?). The predictors were formed from a review of the literature on data reuse as well as an analysis of interviews with actual data reusers. After establishing an effective model for predicting reuse, I examined the relationship between the predictors and the impact metrics I developed in Chapter 5 using bivariate analysis due to the small sample size of reused studies.

The independent variables were comprised of information about the data, documentation, and data producers. The primary data sources for computing these factors were ICPSR's study metadata as well as the citation information and networks I derived in the prior sections. In addition, I produced new factors using additional network data (for example,

the co-authorship networks of the data producers). The main finding from this section was there were two kinds of data that were reused more than other data. First, big data, especially datasets with many variables, were more likely to be reused and more likely to be reused more often than smaller datasets. Second, active data, data that is frequently published, was more likely to be reused and also more likely to score highly on several impact measures. The table below summarizes my study:

Table 3.1 Study summary

Study Section	Research question	Method	Data sources	Outcome
Part 1: Identifying data reuse in a corpus of social science literature	How and why do social scientists cite data?	Analysis of citations to data in social science literature	10-year corpus of social science literature related to datasets in sample; derived from ICPSR's bibliography of data-related publications	Identification of primary and secondary publications related to data Categorization of citation context for data citations
Part 2: Quantifying the impact of data reuse	Do different methods of measuring reuse impact show different datasets to be high impact?	Bibliometric analysis of data citations	Publications from Part 1; additional citation network information from Scopus	Derivation of 5 metrics for data reuse impact Identification of high- and low-reuse datasets
Part 3: Anticipating data impact	What factors are associated with reuse? What factors are associated with high impact reuse?	Logistic regression; bivariate analysis	Dependent variable: Metrics from Part 2 Independent variables: factors defined based on prior research on data reuse and analysis of reuser interviews	Comprehensive list of factors that may influence data reuse Identification of factors associated with datasets that are high-impact by various metrics

3.2 Study setting and core data

Fundamentally, this dissertation revolved around reuse of a collection of datasets, specifically a subset of the data curated by ICPSR. In Chapter 4, I examined papers that are related to these datasets in order to explore citation patterns in the social sciences; in Chapter 5, I used information about these datasets to develop and test reuse impact metrics; and finally, in Chapter 6, I identified what factors were associated with the level of reuse each of these datasets have received.

My sample was drawn from datasets held at The Inter-university Consortium for Political and Social Research (ICPSR), a data repository housed at the University of Michigan. ICPSR is a major source of social science data, with holdings containing more than 65,000 individual datasets from about 8,000 studies. I focused on ICPSR and its data holdings for several reasons. ICPSR is widely recognized as a leader in data archiving in the social sciences, and as a result, the repository is a trusted source of data for social scientists. I was interested, in part, in identifying instances of reuse, and so this trust was important: social scientists may be less likely to use data provided by an organization that they see as untrustworthy (Yakel, Faniel, Kriesberg, & Yoon, 2013).

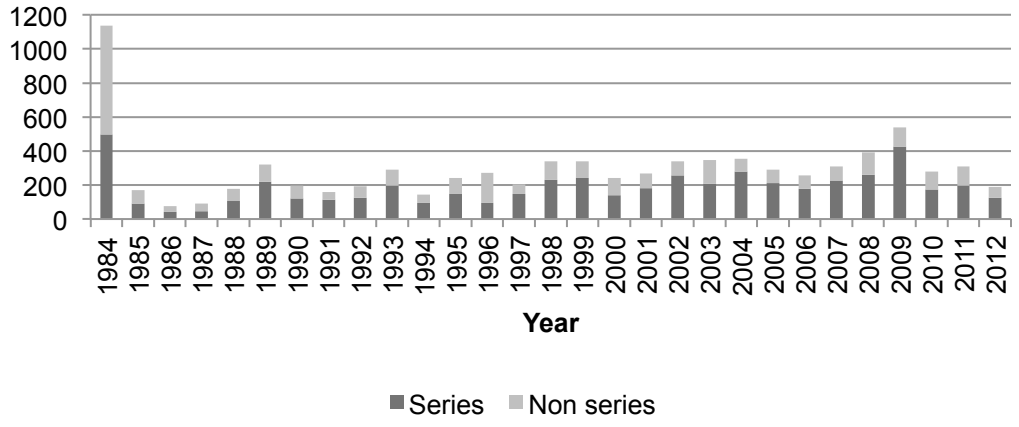
Further, ICPSR curates data from a range of social science disciplines, which allowed me to identify patterns broadly in the social sciences as well as make comparisons between different disciplines within them. Finally, ICPSR was a sensible choice for this dissertation for logistical reasons: I have previously worked with staff at ICPSR and have some familiarity with the organization, and more crucially, ICPSR tracks usage of its data

closely in their Bibliography of Data-Related Literature (described in more detail below), which is an important data source for this dissertation.

3.2.1 Study sample

As of Oct 23, 2012, ICPSR maintained bibliographic records for 8,471 studies. Figure 3.1 includes both those that are part of a series and those that are not and shows the distribution of studies released since 1984. Study release dates were extracted from the *distDate* element in ICPSR's DDI metadata using *minidom*, a Python XML parser.

Figure 3.1 Studies released by ICPSR by year (*N* = 8,471)



Of the 8,471 studies at ICPSR, 5,374 (63%) are series studies. Series studies are those that are ongoing, with new waves of data added periodically. An example is the General Social Survey,[7] which produces a new dataset every two years. Datasets that comprise part of a series like the General Social Survey are excluded from my analysis for two main reasons. First, allocating citations to individual datasets within a series is complicated: users may draw on all extant datasets for their analysis, or they may use just

[7] http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/28

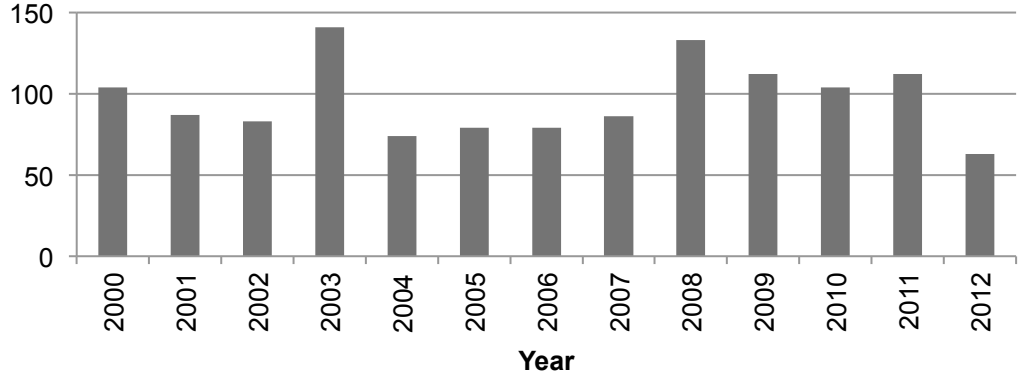
a few. Assigning citations to the series rather than to individual datasets makes sense on the face, but it complicates the analysis here, which explores how datasets gain citations over time. Further, series datasets are different in a crucial way from stand-alone datasets. Series datasets are often designed with reuse in mind. Rather than answering a specific question, they gather broad data that can suit a range of concerns. And each dataset is inextricably linked to the other datasets in the series; while they may not be identical year to year, they are generally very similar. The variation between series and non-series datasets is likely to be much greater than the variation within either group, and for this reason, the two groups must be treated separately.

Excluding series datasets, though, has an important implication for this study. Since series datasets are designed to suit multiple purposes, instead of being tailored to a specific research question the way that smaller datasets often are, they may be more likely to be reused than non-series datasets. Thus, my results in this dissertation are primarily applicable to data produced in the course of smaller projects, or 'long-tail' data. These data, though individually small, are numerous and reuse of such data is not well understood. Further, these data often lack funding or other support for preservation and curation, and are at risk of being lost. Though excluding series datasets will likely eliminate some of the highest-impact data at ICPSR, focusing on non-series data will make an important contribution to understanding reuse of data that is otherwise easily overlooked.

Excluding series datasets leaves 3,097 non-series studies that are eligible for my sample. These 3,097 studies were released at various times during ICPSR's long history: some were originally distributed on magnetic tapes, others as CD-ROMs, and more recently, as digital files. ICPSR instituted direct downloads of data in 2000. Before 2000, data reusers had to request physical copies of the data through their organizational representative; once data became accessible online, end-users could directly access the data themselves and usage increased significantly. Because the time at which data were made available influences how accessible they were—which in turn could influence how much they were reused—I am limiting my sample to the three-year period immediately following ICPSR's institution of direct downloads.

Of the 3,097 total non-series datasets at ICPSR, 1,257 were released in 2000 or later, meaning that they were all accessible by direct download the entire time they have been available. Since 2000, ICPSR has released on average around 327 studies per year, with roughly 70% of those studies part of a series. Figure 3.2 shows studies released by ICPSR since 2000, the first year that direct downloads of data became available.

Figure 3.2 Non-series studies released by ICPSR by year (*N* = 1,257)



My final sample is restricted to non-series studies released between Jan 1, 2000 and Dec 31, 2002. This amounts to 104 studies released in 2000, 87 in 2001, and 83 in 2002, for a total of 274 studies that met my criteria. One study, Minority Aging and Health (ICPSR 3297) was removed from the sample because it has never been available online, just as CD-ROMs, leaving 273 studies in the final sample. The sample studies are listed in Appendix A.

Another important benefit of choosing my sample in this way is that ICPSR began intensively tracking citations in 2000. This means that there are close to 10 years of extensive citation data (2,323 documents, some of which cite more than one dataset) for the datasets released 2000 – 2002, which will allowed me to identify patterns of use over time.

Table 3.2 summarizes my dataset sample selection process:

Table 3.2 Derivation of final study sample (beginning *N* = 8,471, final *N* = 273)

Criteria	Count
Total number of ICPSR studies	8471
- series studies (5374)	*3097*
- studies released before Jan 1 2000 and after Dec 31 2002 (2823)	*274*
- studies not available for direct download	*1*
Final sample	**273**

3.2.2 Extracting study information

Once I identified the sample of datasets that are the focus of this dissertation, I needed to gather information about them, which I accomplished by harvesting and processing the metadata ICPSR has created for the studies. This section describes the process of

extracting information about the studies, and the next section uses this information in a descriptive analysis of my study sample.

I first parsed ICPSR's DDI metadata, which provides extensive information about each study. While some information was extracted as-is from the metadata, I conducted additional processing on most elements. Table 3.3 describes the information gathered from the metadata, the corresponding DDI element, and the details of any additional work.

Table 3.3 Information extracted from DDI study metadata

Information	Corresponding element	Original format (example DDI record)	Processing
StudyID	*IDNo*	<IDNo>9961</IDNo>	*n/a*
Release Date	*distDate*	<distDate date="2000-05-17">2000-05-17</distDate>	*n/a*
Number of funders	*fundAg*	<fundAg>United States Department of Justice. Office of Justice Programs. National Institute of Justice</fundAg>	*Count of number of funding agencies*
Number of authors	*authEnty*	<AuthEnty affiliation="Cornell University">Street, Lloyd</AuthEnty>	*Count of number of authors*
Type of author	*authEnty*	<AuthEnty affiliation="**Cornell University**">**Street, Lloyd**</AuthEnty>	*Extracted raw text; coded for author type*
Processed or unprocessed ?	*topcClass*	<topcClas source="archive" vocab="ICPSR subject classifications">ICPSR XVII.E. Social Institutions and Behavior, Crime and the Criminal Justice System</topcClas> <topcClas source="archive" vocab="NACJD subject classifications">NACJD VII. Crime and Delinquency</topcClas> <topcClas source="archive"	*If owner is Publication-Related Archive, study is unprocessed ; otherwise, study is processed*

		vocab="RCMD subject classifications">RCMD I. Crime</topcClas> <topcClas source="archive" vocab="RCMD subject classifications">RCMD XIII. Race and Ethnicity</topcClas>	
Number of keywords	*keyword*	<keyword vocab="thesaurus">communities</keyword> <keyword vocab="thesaurus">crime</keyword> <keyword vocab="thesaurus">crime statistics</keyword> <keyword vocab="thesaurus">criminal histories</keyword>	*Count of number of keywords*
Extent of geographic coverage	*geogCover*	<geogCover>California</geogCover> <geogCover>Oakland</geogCover> <geogCover>United States</geogCover>	*Count of number of geographic coverage elements*
Time coverage	*timeCvg*	<timePrd event="start" date="1976" cycle="P1"> </timePrd> <timePrd event="end" date="1982" cycle="P1"> </timePrd>	*Calculated number of years covered, if element exists*
Data types	dataKind	<dataKind>administrative records data</dataKind> <dataKind>survey data</dataKind>	Extracted raw text; coded for data type
Data source	dataSrc	<dataSrc>Part 1: Criminal Oriented Records Production System from Alameda County, Part 2: District Attorney Legal Information System, Part 3: Telephone interviews, Parts 4-5: Surveys</dataSrc>	Extracted raw text; coded for data type
Sampling procedure	sampProc	<sampProc>Parts 1 and 2: Not applicable. Part 3: Random sampling. Parts 4-5: Unknown.</sampProc>	Identify whether element exists
Analytic unit	anlyUnit	<anlyUnit>Individuals.</anlyUnit>	Identify whether element exists
Archiving	cleanOps	<cleanOps>Standardized missing	Identify

information		values.</cleanOps> <cleanOps>Performed recodes and/or calculated derived variables.</cleanOps> <cleanOps>Checked for undocumented or out-of-range codes.</cleanOps>	whether element exists
Italic text indicates elements that both processed and unprocessed studies may have. Only processed studies contain the elements in plain text.			

Not all the elements listed above are present for every study. A subset of these elements only exist if a study has been processed, by which I mean that the study has been reviewed by ICPSR staff and had additional metadata added. Unprocessed studies are distributed as received from the data producer, with only the metadata provided by the data producer at the time of submission. The elements added during processing are data type, data source, sampling procedure, analytic unit, and archiving information. The other elements are generally present across all studies, though there are missing values for some.

The following sections address each DDI element in turn, describing what the element is and how it is encoded. Additionally, for each element, I describe how the corresponding characteristic is distributed within the sample. Following the discussion of each individual element, I provide a descriptive analysis of the sample overall.

3.2.2.1 Funders

Each study's funder is encoded in a separate *fundAg* element; if there are multiple funders, there are multiple elements. There is no time information associated with the funder's name, and multiple funders can indicate a succession of funding sources throughout the lifetime of a project, or that the project was co-funded from the beginning.

However, only a minority of studies had multiple *fundAg* elements, so the distinction is not likely to be important.

A substantial minority of studies (88, or 32%) did not contain this element. Of the remaining 185, 167 listed a single funder. The maximum number of funders was eight. Because there was relatively little variation within this variable, I recoded this variable into three levels: no information provided (32%), one funder (61%), and multiple funders (7%).

3.2.2.2 Number of authors

Each author of a study is embedded in a separate *authEnty* element, and I produced a count of *authEnty* elements. All studies contained one or more *authEnty* elements, which included both a data producer's name and his or her affiliation.

The majority (145, or 53%) had a single author, 82 had a pair of authors (30%), and the remainder (17%) had between 3 and 7 authors. This count does not distinguish between named individual authors (e.g., Joe Q. Professor) and agencies (Bureau of Justice) that produce data.

3.2.2.3 Type of author

I also extracted author names and affiliations from the *authEnty* elements to a separate file. Using the names extracted to a separate file, I coded the data authors into 5 groups: individuals, groups of individuals, government agencies, institutions, and media organizations. Though it is likely that there are individuals involved beyond those who

are listed as data producers for the study (an issue which will be addressed in Chapter 4), here I focus only on the number and type of credited authors.

In a large majority of the studies, more than 80%, individual or groups of individuals produced the data. Of the 273 studies, 104 had individual authors (38%); 123 had groups of 2 or more researchers (45%); 30 were produced by governmental organizations (11%); 7 by non-governmental institutions (3%); and 9 by media organizations (3%). Among the studies with groups of authors, most (78, or 63%) were produced by pairs of authors. Only 16 (13%) had 4 or more authors, ranging to a maximum of 7.

Given the relatively low prevalence of governmental, institutional, and media organizations as data producers in this sample, they are treated as a single group for the purposes of this study.

3.2.2.4 Processing status

Many studies at ICPSR are 'processed' before they are released to the public. Data processing involves content authentication (making sure that the correct number of files have been received and can be opened, creating an inventory of the data and documentation files, etc.) as well as checks for completeness and any potential confidentiality issues. The different archives at ICPSR have somewhat different processing guidelines, and individual processors also have some leeway in determining to what extent a given collection is processed (Gutmann, Schurer, Donakowski, & Beedham, 2004). The Publication-Related Archive, though, does not process data at all.

Thus, any data belonging to that archive can confidently be coded as unprocessed. Other data are processed, to varying degrees.

topcClass, the element in which the owning archive name is recorded, is a repeatable element, and some studies had more than one owning archive. Of the 52 studies I identified as part of the Publication-Related Archive, 11 had more than one *topcClass* element. To ensure that none of these studies had been processed by one of their other owning archives, I examined their metadata manually. Each contained an access note (which would be displayed to users) specifying that the data was unprocessed and being distributed exactly as it had been received from the producer. This confirms that these studies are unprocessed.

3.2.2.5 Number of keywords

ICPSR displays keywords as subject terms on study information pages, and they link to other studies related to the same topics. The keywords are subject terms such as 'law enforcement,' 'patient care,' or 'crime mapping,' not terms relating to study methodology, data collection processes, etc.

For this element, I produced a count of the number of individual keyword elements; each keyword is encoded in a separate element. The number of keywords ranged from 0 to 50, averaging 8 (M = 8.1; median = 8). I created levels from the distribution of the number of keywords as follows according to its quartiles: 0; 1-5; 6-8; 9-10; and >11. Only 9 studies (3%) had no keywords listed.

3.2.2.6 Geographic coverage

I also counted geographic coverage. The number of geographic coverage elements ranged from 1 to 69, but both the median and mode were 1. A complication with this element is that some of the coverage terms nest and others do not. For example, Study #3491 (Survey of Citizens' Attitudes Toward Community-Oriented Law Enforcement in Alachua County, Florida, 1996) has geographic coverage elements "United States" and "Florida," whereas Study #1216 (State Exports and the Asian Crisis) has elements "Asia," "Global" and "United States." This means that the count of elements is not easily interpreted: more terms is not necessarily indicative of either broader or more specified coverage.

To address both problems, I recoded this variable in three levels: no information present (8%), one coverage term (45%), or more than one (47%).

3.2.2.7 Time coverage

Time coverage is encoded in the study metadata in two ways: either as a single event (e.g., "1998") or a start and end date (e.g. "1998 to 2002"). The value I used for time coverage was derived by counting the number of years covered. Datasets with their coverage encoded as a single event were given the value '1'; a dataset with coverage 1998-1999 would get the value '2'; etc. A potential issue with this approach is that a single year, for example, '1998' could represent either some subset of months in 1998 or the entire year, or a range, like '1998-1999' could mean two full years or some span of months that begins in 1998 and ends in 1999. A better approach might be to use months of coverage rather than years, but that information is not available.

The distribution of time coverage is quite skewed as well. There are 69 studies (both processed and unprocessed) that do not contain time coverage information. Of the rest, 72 studies have single-event time coverage. 132 studies range from 2 years of coverage up to 203. Coverage is not the same as the time over which the study was conducted; rather, it is the time period on which the study focuses. Studies spanning a very long time period tend to be those that are historical in nature. Study #3226 (Homicides in New York City, 1797-1999 [And Various Historical Comparison Sites]), for example, reviews records of homicides in New York City over a period of more than 200 years. I recoded this variable into four levels: no information provided (25%), single-event (26%), 2-4 years (28%), 5 years or more (21%).

3.2.2.8 Data types

This element applies to processed studies only ($N = 221$). In DDI, the *dataKind* element is described as the "type of data" included in the file.

dataKind is a repeatable, free-text field. In some cases, multiple data types were combined into one element (e.g. "survey data, aggregate data, and administrative records data") while other times, each individual data type was broken out into a separate *dataType* element. I extracted each *dataType* element for the studies in my sample and assigned a single code to each study to identify its data type. The codes are listed in Table 3.4. The first 11 data types are taken directly from ICPSR's taxonomy of data types. I added two final codes to indicate studies that used multiple types of data and those that did not specify the kind of data employed.

Table 3.4 Distribution of ICPSR-assigned data types among processed studies (*N* = 221)

Data Types	Number of studies
Survey data	86
Administrative records data	33
Event / transaction	6
Clinical data	3
Aggregate data	2
Census/enumeration data	2
Program source code	2
Medical records	1
Experiment	0
Machine-readable text	0
Observational data	0
Roll call voting data	0
Multiple	66
Unspecified	20

Of the 221 processed studies in the sample (*N* = 273), 86 (39%) contain only survey data. The next largest single data type is administrative data (33 studies, 15%). Together, these two data types represent slightly over half the processed studies (54%). The remaining data types comprise only 7% of the total. A large group, 66 studies, combines one or more data types. Within that group, most (47 studies) use two data types. The numbers drop off sharply beyond that: 15 studies use three data types, and four studies use four data types, the maximum in the sample.

For this study, I consider administrative data broadly to include data recorded for purposes other than a research project or other data gathering effort. Under this definition, several data types can be rolled into administrative data. 'Aggregate data' in these studies includes things like data collected from tax returns or crime reporting data

aggregated at the level of the county or state. 'Event / transaction data' includes arrest records and other case file details. 'Medical records' are produced through the process of medical care, so this data type fits my broader definition of administrative records as well. I also folded 'census/enumeration data' into 'survey data' and eliminated unused codes.

The final distribution follows:

Table 3.5 Distribution of recoded data types among processed studies (*N* = 221)

Data Types	Number of studies
Survey data	92
Administrative records data	46
Clinical data	3
Program source code	2
Multiple	58
Unspecified	20

Recoding eliminates several of the codes with low counts, and reduces the number of studies coded as 'multiple' as well. (A study previous coded 'multiple' because it listed both administrative and event/transaction data is now counted under administrative data instead.) However, the group remains large. The following table shows the various combinations of data types in the sample:

Table 3.6 Distribution of data type combinations among studies using multiple types of data (*N* = 58)

Data Type Combinations	Number of studies
Administrative + survey	39
Administrative + clinical + survey	5
Administrative + observation + survey	5
Survey + clinical	4
Administrative + clinical	2
Administrative + machine text	1
Survey + focus group	1
Survey data, administrative records data, test score data, and clinical data	1

By far, the most common combination of data types is administrative with survey data. Of the 58 studies with multiple data types, 39 (or 67%) combine these two data types. Eleven other studies add a third data type to this combination.

3.2.2.9 Data source

This element applies to processed studies only. Data source is defined in DDI as the element "[u]sed to list the book(s), article(s), serial(s), and/or machine-readable data file(s)—if any—that served as the source(s) of the data collection." However, in most cases, the text provided in this field contains the type of instrument used to collect the data.

Data source is a free-text field, and producers provided a range of responses. Some examples of data sources my sample studies list are "administrative records data from the Virginia Department of Juvenile Justice and the youths' parole officers" (Study #2730) and "interviewer-administered questionnaires, urine analysis drug tests, biological test results of HIV serostatus, and self-administered questionnaires" (Study #3023). Many of

the studies provided somewhat generic source terms, i.e., “telephone interviews” or “mail-back questionnaires.”

I extracted the raw text and coded it to standardize the set of data sources. Table 3.7 shows the codes I employed. To the extent possible, I retained common wording used in the free-text responses. Studies with three or more sources were coded as ‘multiple,’ and studies that either provided text that did not specify their sources or did not have information for this element were coded as ‘unspecified.’

Table 3.7 Distribution of data sources among processed studies (*N* = 221)

Data Source	Number of Studies
Administrative records	40
Interviews	35
Self-administered survey	34
Interviews + administrative	16
Mixed-method survey + administrative	11
Self-administered survey + administrative	9
Mixed-method survey	7
Census	4
Clinical data	2
Survey + administrative	3
Code	2
Existing data	2
Machine-readable text	1
Media	1
Observation	1
Multiple	23
Unspecified	30

The overall distribution mirrors that of the data kind field, as expected. The most common sources of data are surveys, administrative data, and a combination of the two. The data source field provides more detail on the data than the data type field, though.

Roughly equal numbers of studies used self-administered surveys (generally, mail surveys) and interview survey methods (both telephone and in-person). These codes include studies that conducted multiple waves of the same kind of survey. Fewer studies relied on mixed-method surveys (i.e., mail surveys followed up with interviews). When combining surveys with administrative data, though, more studies engaged in interviews or mixed-method surveys than self-administered surveys.

Two studies present an unusual case, reporting that their data sources are existing data. These two studies (3293 and 3410) both created new datasets based on prior data, the National Survey of Families and Households (NSFH), Wave I, 1987-1988 and Wave II, 1992-1994 (also held by ICPSR, Study 6906). These are not series datasets: series datasets add data periodically under the same study ID number, while in these two cases, additional waves are processed and cataloged as separate studies. In both cases, the data producers focused on subsamples of the NSFH, extracting data from a selection of cases, abstracting existing variables, and creating new variables based on the existing data (and in the case of 3410, combining some data with 1990 Census data to create new variables).

3.2.2.10 Sampling procedure

I initially intended to take the same approach to the text provided for the *sampProc* element as I did for the data source and data kind elements, but data producers' usage of the element posed a problem. The individuals who filled out the data deposit forms from which this information is initially gathered interpreted 'Sampling Procedure' in two distinct ways. Some specified their sampling technique, e.g. "random sampling" (Study #2021), "multistage, stratified, probability sample" (Study #2851) or "convenience

sampling" (Study #3414). Others, though, provided a description of their samples, e.g. "The sample comprises 4th-, 6th-, and 8th-grade students attending two public schools in Berkeley, California, and three public schools and two Catholic schools in Oakland, California" (Study #2341) or "The sample of respondents was drawn from lists of registered voters in the St. Louis and Indianapolis metropolitan areas" (Study #2962). Still others provided text that combined both approaches. Given the difficulty in standardizing this wide variation in responses, I decided to simply code sampling procedure information as present or absent.

Of the 221 processed studies, the majority contained sampling information (162 studies, 73%).

3.2.2.11 Unit of analysis and Archiving information

The text provided for unit of analysis varied widely, while the archiving information provided did not vary at all. As a result, both these elements were coded as present or absent.

The majority of processed elements had these elements present. 133 studies (60%) contained unit of analysis information, and 156 (71%) included archiving information.

3.2.2.12 Dataset size: Number of variables, number of cases

An additional piece of information I included is the number of variables in the studies. This is not in the DDI metadata but in a different database maintained by ICPSR and used to support the Social Science Variable Dictionary, an ICPSR resource that cross-links

variables between studies. I created a SQL query to retrieve the number of variables for each dataset in my sample. Only a limited number (44) have the number of variables encoded anywhere, so for the remainder, I downloaded the studies and recorded the number of variables they contained manually. The number of cases in the data is not recorded in study metadata, so for all datasets in the sample, I downloaded and extracted the number of cases manually.

The number of variables varies dramatically, from a minimum of 0 (for a study that contained only code, not actual data) to a maximum of 6427. The mean number of variables is 488, with a median of about 170. The number of cases similarly varies, from 0 to more than half a million. On average, studies have around 10,000 cases, though the median is much lower, 971.

3.2.2.13 Dataset Subject Category

In addition to extracting information about these datasets from their DDI metadata, I also gathered information about the dataset's subject area from other sources. My primary source of additional information about the data and data producers is Scopus, a citation database maintained by Elsevier. (See Chapter 5 for a full justification of my choice of Scopus as a data source.)

In order to determine the discipline to which each dataset belong, I turned to Scopus' Subject Category (SC) definitions. For each journal that they index, Scopus assigns one or more of four top-level categories (Health Sciences, Social Sciences, Physical Sciences, and Life Sciences) and one or more of 27 lower-level classifications (for example,

Psychology, Computer Science, and Dentistry). I initially planned to assign datasets to disciplines based on the SCs of their primary publications: wherever a dataset was first published would define its SC. However, only 129 datasets have primary publications in the ICPSR Bibliography of Data-Related Literature; the remainder cannot be assigned SCs through this method. (See Chapter 4 for a full explanation of how I identified datasets' primary publications.)

While the datasets do not all have primary publications in the bibliography, they all have authors listed, and Scopus assigns SCs to authors as well, based on their publication records. I assigned disciplines to the datasets based in the SCs of their authors; where the authors had more than one SC, I used the union of their top three (by publication frequency) SCs.

This approach was still problematic for datasets not authored by named individuals. There are 46 datasets in my sample that have group or institutional data authors, though only 16 distinct authors are represented within this group. For each of these authors, I attempted to determine who the individual authors were by searching for the datasets' original grant proposals and examining the set of papers citing the data to see if any individual data producers were identified. The datasets and the authors I identified are listed in Table 3.8.

Table 3.8 Identification of individual authors for institutional- or government-produced studies

Listed Data Author	Identified Data Producers
European Values Study Group and World Values Survey Association	None
National Center for State Courts	Rottman, David; Flango, Carol; Cantrell, Melissa; Hansen, Randall; LaFountain, Neil: Authors listed in "State Court Organization 1998"; Langton, Lynn; Cohen, Thomas: Authors listed in "State Court Organization 1987-2004"
National Gambling Impact Study Commission	Gerstein, Dean; Murphy, Sally; Toce, Marianna; Hoffman, John; Palmer, Amanda; Johnson, Robert; Larison, Cindy; Chuchro, Lucian; Buie, Tracy; Engelman, Laszio; Hill, Mary Ann: Authors listed in "Gambling Behavior and Impact Study: Report to the National Gambling Impact Study Commission" (1999)
Pennsylvania Commission on Sentencing	None
Pennsylvania Commission on Sentencing	None
United States Department of Health and Human Services. National Institutes of Health. National Institute on Drug Abuse', '	Fisher, Dennis G.; Rhodes, Fen; Watters, John; Bluthenthal, Ricky; Weatherby, Norman; McCoy, Clyde B.; Anderson, Marcia; Cottler, Linda B.; Robles, Rafaela R.; Williams, Mark; Montoya, Isaac; Desmond, David
United States Department of Health and Human Services. National Institutes of Health. National Institute on Drug Abuse	Hubbard, Robert; Simpson, D. Dwayne; Anglin, M. Douglas; Fletcher, Bennett: PIs listed at www.datos.org/aboutdatos.html
United States Department of Health and Human Services. National Institutes of Health. National Institute on Drug Abuse	None
United States Department of Health and Human Services. Social Security Administration	Krute, Aaron; Burdette, Mary Ellen: Authors listed in "1972 Survey of Disabled and Nondisabled Adults: Chronic Disease, Injury, and Work Disability"
United States Department of Health and Human Services. Substance Abuse and Mental	Gerstein, Dean; Datta, A. Rupa; Ingels, Julia; Johnson, Robert; Rasinski, Kenneth; Schildhaus, Sam; Talley, Kristine; Jordan, Kathleen; Phillips, Dane; Anderson,

Health Services Administration. Center for Substance Abuse Treatment	Donald; Condelli, Ward; Collins, James: Authors listed in "National Treatment Improvement Evaluation Study, Final Report" (1997)
United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Office of Applied Studies	Gerstein, Dean; Schildhaus, Sam: Authors listed in "Services Research Outcomes Study" (1998)
United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Office of Applied Studies	Horgan, Constance; Levine, Helen; Reif, Sharon; Ritter, Grant; Lee, Margaret; Marsden, Mary Ellen; Pieroni, Maria: Authors listed in "Alcohol and Drug Services Study (ADSS): The National Substance Abuse Treatment System: Facilities, Clients, Services, and Staffing" (2003)
United States Department of Justice. Bureau of Justice Statistics	Steadman, Greg: Author listed in "Survey of DNA Crime Laboratories, 1998"
United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics	Langan, Patrick; Levin, David: Authors listed in "Recidivism of Prisoners Released in 1994"
United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics	DeFrances, Carol: Author listed in "State-Funded Indigent Defense Services, 1999"

In the end, I was able to assign discipline via author SC for 253 datasets, or 93% of the sample. The remainder was datasets with institutional authors for which I could not identify individual names (4 studies, 1%) and studies whose authors are not listed in Scopus (15 studies, 5%).

Table 3.9 Distribution of datasets (*N* = 253) across subject categories

Subject Category	Number of Datasets
Medicine + Psychology + Social Sciences	77
Social Sciences	62
Social Sciences + Business, Management, and Accounting + Economics,	15

Econometrics, and Finance	
Business, Management, and Accounting + Economics, Econometrics, and Finance	15
Medicine + Social Sciences	9
Medicine + Nursing + Social Sciences	8
Economics, Econometrics, and Finance + Business, Management, and Accounting + Mathematics	7
Economics, Econometrics, and Finance	5
Medicine	5
Social Sciences + Economics, Econometrics, and Finance	4
Medicine + Psychology	4
Economics, Econometrics, and Finance + Arts and Humanities + Social Sciences	4
Medicine + Neuroscience + Psychology	3
Arts and Humanities + Social Sciences + Psychology	2
Psychology + Social Sciences	2
Health Professions + Nursing + Social Sciences	2
Medicine + Biochemistry, Genetics and Molecular Biology + Social Sciences	2
Medicine + Health Professions + Nursing	2
Social Sciences + Business, Management, and Accounting	2
Biochemistry, Genetics and Molecular Biology + Medicine + Psychology	2
Arts and Humanities + Economics, Econometrics, and Finance	2
Social Sciences + Mathematics + Decision Sciences	1
Social Sciences + Arts and Humanities + Business, Management, and Accounting	1
Economics, Econometrics, and Finance + Mathematics + Social Sciences	1
Social Sciences + Computer Science	1
Social Sciences + Mathematics	1
Social Sciences + Environmental Science	1
Social Sciences + Medicine + Agricultural and Biological Sciences	1
Decision Sciences + Medicine + Social Sciences	1
Social Sciences + Medicine + Computer Science	1
Mathematics + Psychology + Social Sciences	1
Psychology	1
Medicine + Nursing + Psychology	1
Business, Management, and Accounting + Decision Sciences + Social Sciences	1
Social Sciences + Business, Management, and Accounting + Psychology	1
Social Sciences + Arts and Humanities + Medicine	1
Psychology + Health Professions + Social Sciences	1
Medicine + Social Sciences + Economics, Econometrics, and Finance	1
Social Sciences + Environmental Science + Medicine	1
Medicine + Nursing	1

The most common dataset SC was Medicine + Psychology + Social Sciences[8], with 28% of the studies falling into this group. Second most common was the general 'Social Sciences' category (23%) followed by two categories relating to economics and business fields. Of the 40 different dataset SCs, 27 (68%) contained 'Social Sciences' as one of their component subject areas. Nearly half (48%) included Medicine, Nursing or Health Professions as at least one of their component SCs. About a third (33%) contained Business, Management, and Accounting and/or Economics, Econometrics, and Finance. Psychology, another common field, was present in about 28% of the SCs. In addition to the SCs assigned above, I also coded datasets based on whether 'Social Sciences' was their only SC, one of multiple SCs, or not an SC they belonged to at all (Table 3.10).

Table 3.10 Distribution of datasets (*N* = 312) within and outside of social sciences

Subject Category	Number of datasets
Social Sciences only	62
Social Sciences + another SC	143
Another SC only	48

3.2.3 The sample data

After extracting and assembling the study information described above, I examined the characteristics of my sample with two major goals in mind. First, I wanted to determine whether the yearly cohorts of data were statistically similar to one another. Second, I sought to find out if there were differences between processed and unprocessed studies beyond the information added through processing. I primarily used chi-square tests to

[8] 'Social Sciences' is the only category label assigned at both the top and lower levels.

compare populations within my sample, and Fisher's exact test in cases where counts in any level of a variable were low (less than five). All statistics throughout this dissertation were computed using Stata 12 (StataCorp, 2011).

The most common variety of study in the sample is a survey produced by an individual author (37 studies; 14%), followed by surveys produced by groups of authors (33 studies; 12%). Also common are studies by individuals and groups that combine two or more multiple data types (52 studies; 19%).

I compared studies in the sample on the variables both processed and unprocessed studies share: number of funders ($X^2(4, N = 273) = 1.86$, $p = 0.76$), author type ($X^2(4, N = 273) = 2.37$, $p = 0.66$), processing status ($X^2(2, N = 273) = 0.57$, $p = 0.75$), number of keywords ($p = 0.91$, two-tailed Fisher's exact test), geographic coverage ($X^2(4, N = 273) = 4.14$, $p = 0.39$), time coverage ($X^2(6, N = 273) = 6.56$, $p = 0.36$), number of variables ($p = 0.97$, two-tailed Fisher's exact test), and SC group (Social Sciences only, Social Sciences plus another field, another field only) ($X^2(4, N = 253) = 4.76$, $p = 0.31$). There were no significant differences between the yearly cohorts on any variable.

I tested whether studies in different SC groups differed in their processing status and found that studies that were affiliated only with fields other than Social Sciences were less likely to be processed ($X^2(2, N = 253) = 39.84$, $p < 0.001$). There were a number of significant differences between unprocessed and unprocessed studies. Unsurprisingly, processed studies tended to have more complete and more detailed documentation.

Unprocessed studies in this sample were less likely to have funder information provided ($p < 0.001$, two-tailed Fisher's exact test). Further, they were less likely to have more than 5 keywords ($p < 0.001$, two-tailed Fisher's exact test), more than 1 geographic coverage term ($p < 0.001$, two-tailed Fisher's exact test), or any time coverage terms ($X^2(3, N = 273) = 121.43, p < 0.001$). No unprocessed studies had a count of the number of variables included in the study. Finally, no unprocessed study was produced by a government agency, institution, or other organization. This suggests that there is a relationship between author type and whether or not ICPSR processes a study ($X^2(2, N = 273) = 13.02, p = 0.001$). Studies produced by government agencies, institutions or other organizations are often specifically funded for curation by ICPSR; these organizations contract with ICPSR to process the data and make them available, so this relationship is not surprising.

I also divided the data into two subsets, processed and unprocessed studies, and conducted comparisons within these groups to determine if they are uniform year-to-year. For the unprocessed group, I examined the same set of variables as in the previous analysis. For the processed group, I included the set of variables that are added when a study is processed: data type and number of data types; data source; sampling procedure; unit of analysis; and archiving information. There were no significant differences year-to-year among either the unprocessed data or the processed data on any variable.

This analysis demonstrates two things: first, there are measurable differences between processed and unprocessed studies, and processed studies have more complete

documentation than unprocessed studies, which shows that processing status can be used as a proxy for documentation quality. Second, the differences arising from processing status, along with the differences in author type, which appear to covary with processing status, are the only differences within the sample. There are no significant differences along any axis between the yearly cohorts I have selected into my sample, which means that there were no changes in the types of data released year to year or how the data were documented.

Together, these findings suggest that this sample is robust enough to build my dissertation study upon. The 273 studies that are part of my sample represent a complete census of non-series data released between 2000 and 2002 from ICPSR. Based on the findings above, this three-year sample can be treated as a uniform cohort rather than as three independent groups.

3.3 Limitations

There are several important limitations to my study. One key issue is that I am focusing on data from ICPSR. While ICPSR is an important data source for the social sciences, it is not the only source, and it may be the case that the data held there and the users of that archive are not representative of the social sciences as a whole, limiting the generalizability of my findings.

Other limitations are specific to the design of individual sections of the dissertation. I address issues relating to statistical reliability individually in each chapter.

The analysis in Chapter 4 relies heavily on the ICPSR data-related publications bibliography. Though this resource is carefully compiled, it is not error free and it is impossible to know if it is comprehensive. In fact, it is likely that there are data-related publications in the world that are not captured in the bibliography, either because they did not cite the data clearly or through simple oversight. There is extra staff dedicated to collecting citation information relating to the National Archive of Criminal Justice Data's holding, which means that those studies may be overrepresented in the bibliography, which in turn means that the particular citation habits in that domain may skew my findings overall.

In Chapter 5, I used Scopus to compile bibliometric networks of several types. No single data source has complete coverage of all literature, and all contain occasional errors, like a single publication appearing more than one time due to mistakes in the title, for example. Thus the citation counts have a certain amount of error in them, and so do the bibliometric networks, as publications that are not in the database will not appear, and others may be duplicates.

Identifying the factors which are associated with data reuse and impact in Chapter 6 is dependent, in part, on first putting together a comprehensive set of candidate factors. In this, I am limited in a few different ways. Most pertinently, only factors that can be measured can be included. Professors mentioning data in classes might influence students to use those data, but 'word-of-mouth' cannot be included as a factor because I cannot effectively quantify the amount of word-of-mouth each dataset receives. Additionally,

even those factors that can be measured may be an approximation. For example, I can use an individual's h-index as a proxy for their reputation, but this is not an exact match.

Despite these limitations, this dissertation study will present important findings related to measuring and anticipating data reuse impact. This study represents a critical exploratory first step towards developing better methods for identifying data reuse and quantifying its impact as well as using the knowledge gained from those methods to inform curation and funding decisions.

CHAPTER 4

Data Citation Patterns in the Social Sciences

Data are fundamental to research, and the social sciences use data in a variety of ways. Some papers document the collection and analysis of new data; others describe secondary analyses or reanalyses of existing data. Still other publications turn a critical eye toward prior studies, or synthesize a body of work to develop new ideas. While there are standard practices for citing publications in new work, standards for citing data are limited or nonexistent. This chapter reports an analysis of a set of social science publications that use data based on ICPSR's bibliography of data-related literature,[9] I examine three categories of publications—primary, secondary, and non-research publications—to understand how and why social scientists cite data in their work.

My work extends Mooney's (2011) work on the ICPSR Bibliography of data related literature by examining a much larger sample of the bibliography. Although ICPSR provides formal citation text to all data reusers and encourages data citation, the question remains: do social scientists cite data? More specifically, do social scientists cite data when they use them, and do they only cite data when they use them (versus citing data in other context, like a literature review)? I also add another dimension to the question of

[9] http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/

data citation. Do data producers use data citations (for example, to let readers know that the data are available) or do only reusers cite data?

4.1 Methods for identifying reuse

Previous studies of data reuse have relied on both automated and manual methods for identifying secondary papers. Sieber and Trumbo (1995) and Mooney (2011) both relied on pre-existing corpora of data-related papers: a bibliography of papers relating to the General Social Survey and the ICPSR Bibliography of Data-Related Literature, respectively. Sieber and Trumbo sampled all 198 relevant, English-language articles and manually coded each for citation elements and locations. Mooney refined ICPSR's bibliography to the subset of articles published by researchers at a single large public university. After defining her sample of 49 articles, Mooney conducted a manual content analysis.

Pienta, Alter, and Lyle (2010) similarly worked from the ICPSR Bibliography of Data-Related Literature but used an automated method of distinguishing between primary publications and others. They developed a script to compare the names of data producers and study team members to the authors of each article, identifying secondary articles as those where none of the research team members are authors.

Brown (2003) and Piwowar et al. (2011), both exploring reuse in genomics, began by compiling collections of data-related literature. Brown gathered a set of literature by searching ISI Web of Science for references to GenBank accession numbers. Piwowar et al. extended this method, searching not only ISI Web of Science but Google Scholar and

other databases, and included searches for DOIs and other unique identifiers in addition to the GenBank accession number. Piwowar et al. also located the data collection article (the original article by the data producers describing the data) for each dataset in their sample and compiled the set of papers that cite that article. Finally, they conducted a manual review of the full text of all articles to determine if the cited data had actually been reused.

In this study, I used a combined approach. Using the ICPSR Bibliography of Data-Related Literature as my literature source, I first identified papers written by data producers and members of study teams as listed in ICPSR's study description metadata, which left a set of reuse candidate articles. Then, I manually reviewed each candidate item to assess whether it was an instance of reuse or another type of scholarly work, such as a literature review or commentary. The process is described in full in the next section.

4.2 Processing the data-related publications bibliography

The bibliography of data-related literature includes reference information about publications related to ICPSR's data. These documents include journal articles, conference papers, reports, theses, and other materials. About two-thirds of the bibliography is comprised of journal articles, with reports forming the next largest group. The remaining categories each form 5% or less of the total. In each document entry, there is a list of the ID numbers corresponding to the datasets to which the document is related.

This information is all encoded in a SQL database, which I was given access to by ICPSR. Citations in the bibliography are connected to their related studies via the study

ID. Using the set of study IDs corresponding to the datasets in my sample, I created a query to retrieve all documents that contained one or more studies in the sample. As of December 11, 2012, there were 2,323 items matching these criteria. Table 4.1 breaks out the documents by type:

Table 4.1 Distribution of documents in Bibliography of Data-Related Literature by type (*N* = 2,323)

Document Type	Number	Percent
Audiovisual Material	3	0.13%
Book	37	1.60%
Book Section	120	5.18%
Conference Proceedings	91	3.93%
Document	6	0.26%
Electronic Source	2	0.09%
Grant report	1	0.04%
Journal Article	1560	67.30%
Magazine Article	2	0.09%
Newspaper Article	2	0.09%
Report	374	16.13%
Thesis	120	5.18%

I used an iterative process to identify secondary publications (publications by authors other than the data producers) and primary publications (publications by the data producers). The first step was automated. I wrote a Python script to compare the last names of the data producers listed for each study to the last names of the authors listed for each publication. If any of the data producers appeared in the publication's author list, the publication was marked as a primary publication; if the publication authors were distinct from the data producers, the publication was labeled as a secondary publication.

Simply comparing last names left several problems unaddressed, however. First, the bibliography does not use a controlled vocabulary for names, so some names are misspelled or appear with some variation, especially those that contain special characters. Second, two different people might share a last name, or the same person may have changed names over time. Finally, some individuals, especially graduate students, may have had a hand in producing the data but were not listed as data producers.

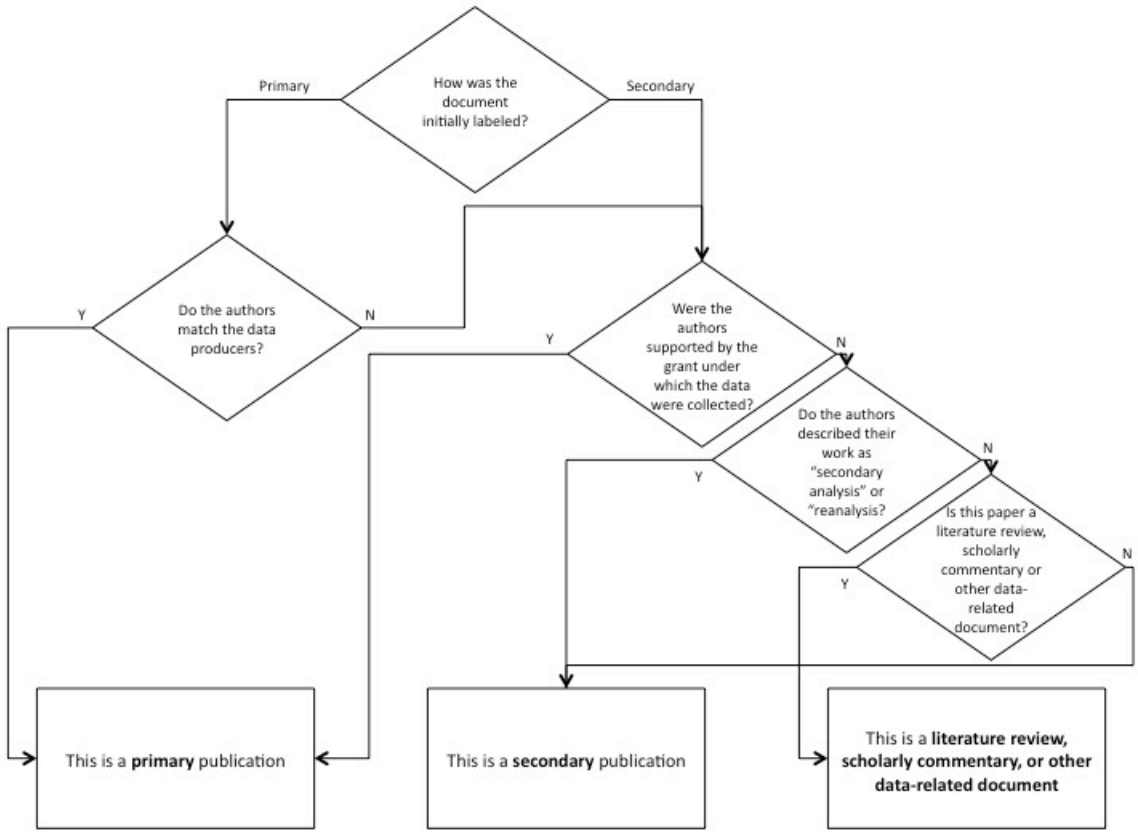
In order to address these complications and validate the labels applied automatically by the name comparison script, I manually reviewed each publication in the sample of 2,323. I acquired the full text of these publications several ways. Most were available online, either freely or through the University Library. A smaller proportion was not available electronically, so I consulted paper versions held by the library. Copies of articles and books not available through the library were requested through inter-library loan. In the cases ($N = 15$) where I was unable to procure full-text versions of a document because the University did not own a copy and inter-library loan could not acquire one, I excluded the document from the sample.

To confirm that a publication was a primary paper, I double-checked the author names to ensure that they matched the data producer names. Since I suspected that some of the secondary publications had been written by authors who had been involved in data collection—albeit uncredited—I also examined the acknowledgements. If the paper was produced under the same grant as the dataset, I counted that paper as a primary publication. One example of this is Grella et al. (2003), which uses the DATOS dataset

(ICPSR #2258). None of the listed authors are credited anywhere as data producers, but the acknowledgements to their paper show that their work was supported by the DATOS grant. Grella and her co-authors were, at the time, located at an institution that hosted a DATOS Research Center, further supporting the case that they were likely involved in primary data collection to some extent.

Further, I was not certain that all papers initially labeled 'secondary' actually represented reuse. It was possible that some papers cited data for other purposes, without actually reusing the data: in a literature review or a summary, for example. In order to distinguish secondary use from other kinds of use, I first examined each abstract. If the abstract specified that the authors conducted "secondary analysis" or "reanalysis" of the cited data, I counted that paper as secondary reuse. If the abstract was unclear about what data were used and how, I read the rest of the paper, focusing on methods sections, particularly the 'data' subsection included in many papers. The following figure illustrates the classification decision tree:

Figure 4.1 Decision tree for classifying publications



I developed a set of codes to indicate the kind of usage each paper represented. These differ from the categories ICPSR applies in that those categories are document types (e.g. journal article, report, etc.), whereas the categories I use have more to do with the content (e.g. secondary use, scholarly commentary, etc.). The categories corresponding to secondary use were secondary analysis / reanalysis (including secondary analysis of a complete dataset or a subset, and combining the data with data from other sources) and use of the data to validate or demonstrate a new method or instrument. Non-reuse categories were scholarly commentary (including letters to the editor, editorials, speeches and addresses, and essays); media commentary and publicity (e.g. press releases and

news articles); literature reviews and meta-analyses of published results; policy documents based on published results; and other non-reuse articles, such as appendices, corrigenda, and bibliographies. Some publications were also excluded from the sample if they did not appear to be related to the data they were supposed to cite or if they were duplicates of other documents in the bibliography.

In the end, after removing duplicates, unrelated publications and publications for which I could not find the full text ($N = 150$), I compiled a sample of 2,173 documents in the bibliography. Table 4.2 shows the distribution of publications across the codes I applied:

Table 4.2 Distribution of documents in Bibliography of Data-Related Literature, after coding and elimination of duplicates ($N = 2,173$)

Document Type	Number	Percent
Primary	1,325	57.04%
Secondary	577	24.84%
Literature review	117	5.04%
Scholarly Commentary	81	3.49%
Other data-related documents	61	2.63%
Media commentary and publicity	12	0.52%

4.3 Identifying data citations and acknowledgements in published literature

After finalizing the set of publications related to the data in my sample, I identified all instances in which authors cited or otherwise acknowledged the data that they used. I grouped the publications into three categories: primary publications written by data producers; secondary publications written by authors other than the data producers; and non-research papers. Non-research papers included literature reviews, scholarly commentaries, and other data-related documents. I consider these papers to be 'non-

research' because they are not hypothesis-driven and generally, while these publications are *about* data and the studies that produced the data and their conclusions, their authors typically do not engage with the data directly.

My analysis of the bibliography used the categorization of papers derived above to answer three important questions. First, using the set of primary publications, I addressed the question of whether or not data producers indicated that their data were or would be publicly available. All the data used in these publications became publicly available between 2000 and 2002; while some of the papers were written before the data were available, many were not. Did data producers say that their data would be available in publications prior to its release data? When data producers wrote papers after data were released and publicly available, did they cite or acknowledge ICPSR or otherwise indicate that the data were available?

The second and third questions are closely related to one another. When social scientists used data that they did not create themselves, did they cite or acknowledge it, and if so, how? And when social scientists wrote papers that were related to data but did not actually use data, did they cite or acknowledge the data? To answer these questions, I focused on the set of secondary papers (papers that use data not created by the authors) plus literature reviews, scholarly commentaries, and other data-related documents (papers related but not using data).

For this analysis, I limited my scope to journal articles only. Journal articles made up 67% (1,473 documents) of the 2,173 items in the bibliography (after removing duplicates, unrelated items and items that could not be located), distributed across content types as follows:

Table 4.3 Distribution of journal articles (*N* = 1,473) by type

Document Type	Number	Percent
Primary	848	57.8%
Secondary	450	30.7%
Literature review	67	4.6%
Scholarly Commentary	79	5.4%
Other data-related documents	23	1.6%

I began coding using a simple system, recording for each document whether there was a data citation/acknowledgement or not. If there was an acknowledgement present, I recorded the form that it took. After completing an initial coding of the first 100 documents, I reviewed the information I recorded and developed the following framework for identifying data acknowledgements. After returning to the initial set of 100 documents to apply the new codes, I completed coding the remaining articles.

4.3.1 Framework for categorizing data citations or acknowledgements

When citing or acknowledging data, authors have two primary entities that they can credit: the producers of the data and the providers. Data producers are the researchers that created the data; the data provider is the people or institution from whom the author actually acquired the data. Sometimes the producer and provider are one in the same, but other times they differ. The papers I examined demonstrated several different approaches to citing data and data producers.

One approach to acknowledging data directly highlighted the original data producer. Some authors provided a formal citation to publications or other documentation produced by the data producer. The contexts of these citations varied. Citations to data producers' work could be an explicit reference to the source of the data used in the analysis, as in Tolbert and Smith (2005, p. 285):

> *Using aggregate-level voter eligible population (VEP) turnout data from the 50 states (McDonald & Popkin, 2001),* we examine whether states that use the initiative process have higher levels of voter turnout in midterm and presidential elections than states that do not permit the process.

Credit for producing the data is clearly assigned here: these data come from these individuals. But a citation to a data producer's work could also serve as a pointer to more information about the data, as illustrated in Reyes-Ortiz et al. (2008, p. 480):

> Data are from the Hispanic EPESE, a population-based study of 3050 noninstitutionalized Mexican Americans 65 years old or older (83% response rate) residing in five Southwestern states: Texas, California, New Mexico, Colorado, and Arizona. *Sampling and data collection are described elsewhere (13).*

In contrast to the first example, where the citation identifies the original data producers directly, this citation identifies them only indirectly. Instead, the primary purpose of this citation is to direct the reader to methodological detail specified elsewhere.

I drew a distinction between citations to data producers used to acknowledge data or provide additional information about data and citations to data producers used as background knowledge. Tolbert and Smith, in the example above, cite McDonald & Popkin (2001) in relation to the data they use in their analysis. However, at other points

in their paper, they also cite papers by both McDonald (p.284) and Popkin (p. 304) separately. The difference between these citations is in their intent. In the first instance above, the citation indicates that *McDonald & Popkin* produced the data Tolbert and Smith use in their analysis. In the citations to McDonald and to Popkin elsewhere, Tolbert and Smith employ more standard literature citations. They cite McDonald and Popkin to provide background on previous studies and to support a claim, respectively.

Some authors acknowledge the producers of the data they use informally, rather than through a formal citation. Mojtabai (2005, p. 340), for example, credits the principal investigator who conducted the survey on which Mojtabai's paper relies on directly in the text:

> Data are from the ''Six State Survey of Elderly Dual Enrollees in Medicare and Medicaid'' *(Principal Investigator: Judith A. Kasper, Ph.D.),* which comprised an age-stratified random sample of community-dwelling adults who were 65 years old or older at the time of interview in 1999 and were enrolled in both Medicare and Medicaid.

Other forms of informal acknowledgement included mentioning the data producer in a footnote or in the acknowledgements, as in Osgood and Anderson: "The authors are especially grateful to *Finn Esbensen for access to these data*" (2004, p. 519).

Authors incorporated another type of acknowledgement into their work as well, by informally or formally citing the data provider. All the data cited in these papers are available from ICPSR, though not every author acquired the data from the repository. Some authors provided formal data citations as part of their reference list, in a footnote,

or in the acknowledgements section. Mojtabai and Zivin (2003) included, for example, the following text in their references list:

> Substance Abuse and Mental Health Services Administration. 2001. ‘‘Uniform Facility Data Set (UFDS) 1999’’ [accessed on April 28, 2002]. Available at http://wwwdasis.samhsa.gov/99ufds/ufds99.htm.

There were also informal methods of acknowledging data and their source, especially by listing a URL from which the data can be acquired, either directly in the text, as in Agosti, Nunes, and Levin (2002, p. 645), or in footnotes or acknowledgements.

The following matrix summarizes the types of data acknowledgements social scientists used in their papers, categorized along two axes: the formality of the acknowledgement and its target.

Figure 4.2 Matrix of data citation types

		Citation style	
		Formal	Informal
Cited entity	Data producer	Citation to previous work by data producer (does not include citations that are primarily part of literature review)	Unstructured identification of data producer in text, footnote or acknowledgements
	Data provider	Citation to data source, typically a repository	Unstructured identification of data source

These two varieties of acknowledgements are not mutually exclusive, of course: a paper may have cited the data producer’s previous publications as well as provided a formal citation to acknowledge the repository that provided the data, for example.

4.4 Findings

In this section, I present my findings in three parts. First, I focus on the set of journal articles that are primary publications and show that primary authors rarely cite or acknowledge their data's public availability in their papers. Second, I turn to secondary publications and non-research publications. In the case of secondary publications, I find that authors most commonly cite data via formal citation to papers written by the data producer, rather than via a citation, formal or informal, to the data's source. However, the occurrence of formal data citations in secondary publications increased over time. In non-research publications, authors almost never cite data directly, nearly exclusively citing data producer publications.

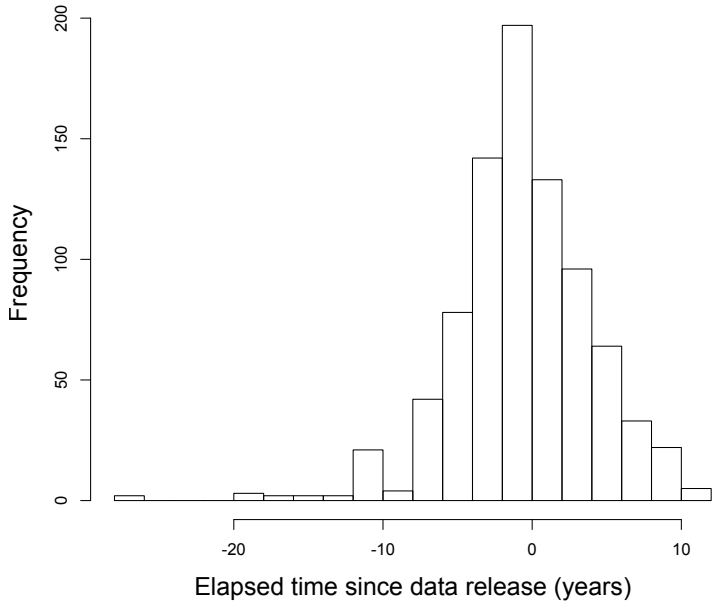
4.4.1 Do authors cite or acknowledge data in primary papers?

Within my sample of 1,473 journal articles related to studies released between 2000 and 2002, I found 848 primary publications (publications actually using the data, written by the data producers). About 60% of the primary papers' publication dates precede the ICPSR release of the data used in those papers (495 papers). Of the remainder, 279 (33%) were published after the data were released, and 74 (9%) appeared in the same year (Figure 4.3).

Two studies, Three-Wave Political Socialization Panel Survey of Children in the San Francisco East Bay Area, 1968-1969 (ICPSR #2341) and Beliefs About Social Stratification, 1980: [United States] (ICPSR #8702) have papers produced 15 years or more before the data were released. In both these cases, the studies were older. Study #2341 was conducted in 1968, and the data producers published from the data as early as

1973. Study #8702 was created in 1980, and papers appeared starting in 1983. These studies were not available online until 2000 and 2001, respectively, giving rise to the long time elapsed between primary publication and data release.

Figure 4.3 Histogram of elapsed time between publication of primary papers and data release. Time is recorded in years, with negative numbers indicating publication prior to data release and positive numbers indicating publication after release.



The median elapsed time between data release and publication of primary papers is 0 years. While there is a relatively wide spread in publication times, the majority of the primary papers ($N = 787$, 93%) were published within three years before or after data release.

Authors rarely mentioned data availability in this set of primary papers. Only 61 papers (7%) provided any type of notification that the data were currently available either from the author or a repository. While 59 of those papers were clear about where the data could be accessed (either ICPSR, another repository, or a website), two papers included

only a passing mention of the data's availability without any further details. Even rarer were indications that the data would become publicly available in the future. Just two papers included statements noting that the data would be deposited at some point after the publication of the paper. Among the 61 primary papers in which data producers acknowledged the availability of their data in any way, either current or prospective, about half (32 papers) cited the same dataset, the National Comorbidity Study (ICPSR #6693). The remaining papers cite 19 datasets, with a median and mode number of citations per dataset of 1.

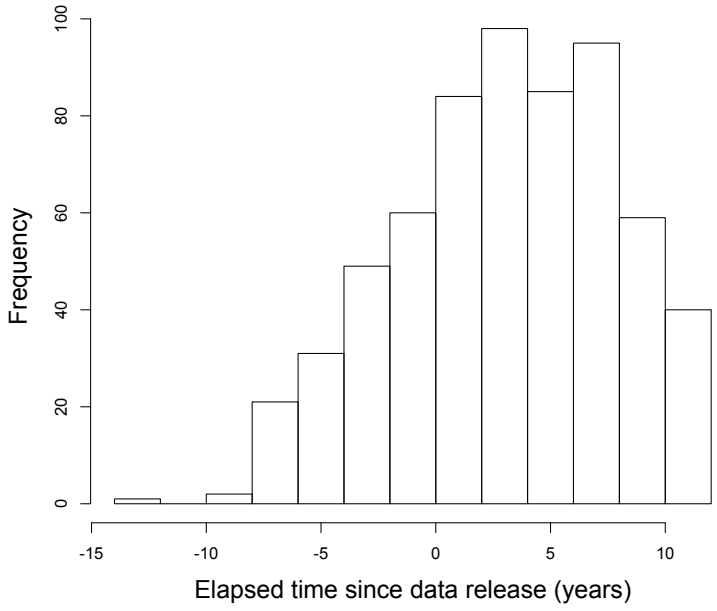
A Fisher test showed that papers published after data were released are no significantly more likely to have some kind of data acknowledgement than those published before data release ($p = 0.55$, two-tailed Fisher's exact test). However, removing the National Comorbidity Study from the sample changes the result. For all other datasets, primary papers published after their release are significantly more likely to provide a data citation, either formal or informal ($X^2(1, N = 696) = 9.09$; $p = 0.002$).

4.4.2 Do secondary authors cite data when they use data? Do authors cite data in documents where they do not use data?

Of the 1,473 total journal articles, I identified 625 non-primary publications in this dataset: 449 secondary publications, 80 scholarly commentaries, 71 literature reviews, and 25 other data-related documents. In contrast to the primary publications, most of these documents were published after the data were released. The majority, 427 publications (68%), were published after the data became publicly available through ICPSR, while 34 articles (5%) were published in the same year the data were released,

and 164 prior (26%). Figure 4.4 shows the distribution of the time elapsed between data release and publication of an article. (Negative times indicate articles published before data became available through ICPSR.)

Figure 4.4 Histogram of time elapsed (in years) between data release and non-primary article publication (*N* = 625). Time is recorded in years, with negative numbers indicating publication prior to data release and positive numbers indicating publication after release.

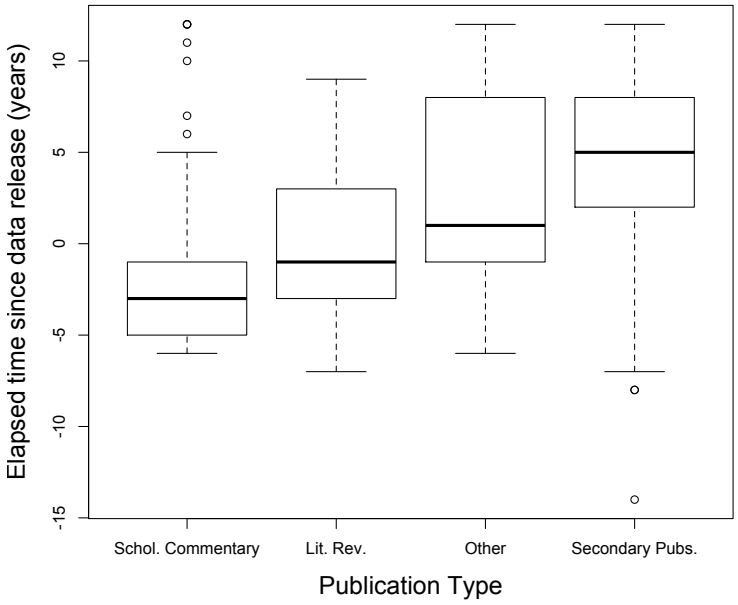


Overall, the median time elapsed between data release and article publication in this sample is 4 years. The earliest any non-primary publication was published was 14 years prior to data release. Of the 43 papers published 5 years or more before data release, 25 are related to a single dataset, the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment (SUPPORT) and Hospitalized Elderly Longitudinal Project (HELP), 1989-1997 (ICPSR #2957), and these 25 papers are all works of scholarly commentary published in 1995 and 1996, as the project was coming to a close. The SUPPORT project is unusual in that it was a very large, federally funded project that

did not produce the expected results: the interventions the project tested were shown to have no effect (Berwick, 1995). This resulted in a flurry of commentary as the project ended and the data producers and other stakeholders attempted to make sense of these outcomes through editorials (with titles like "What SUPPORT Really Means" (Markson & Clark, 1996) and "The Best Laid Plans" (Moskowitz & Nelson, 1995)), and critical letters (Donnelly, 1996; Frank, 1996; Prendergast, 1996).

The histograms below show the distributions for secondary publications and other publications separated.

Figure 4.5 Elapsed time between data release and publication for literature reviews, scholarly commentaries, other data-related documents and secondary publications



A Kruskal-Wallis test showed a statistically significant difference in the elapsed time between data release and publication for the different publication types (Kruskal-Wallis $X^2(3, N = 625) = 163.18$; $p < 0.001$). The median time between data release and

publication for literature reviews, scholarly commentaries, and other data-related documents is one year prior to data release, with a mean of -0.31. On average, 'other' documents are published slightly after data release. This is unsurprising, since this category includes documents like press releases and promotional materials that would be released at roughly the same time that the data become available. Literature reviews appear, on average, in the year prior to data release. Scholarly commentaries also tend to be published before data are released, but this particular sample is somewhat skewed, due to the large proportion of commentary articles related to the SUPPORT study.

Secondary papers are published later than the overall sample, with a mean and median time between data release and paper publication of 5 years. Just 64 articles (16%) were published before the data they use were available through ICPSR. Though some of the data may have been available through other avenues prior to ICPSR's release, the sharp increase in number of articles produced after data were made publicly available lends some initial credence to the suggestion that publishing data through an archive increases their usage.

4.4.2.1 Secondary publications

Turning now to the question of how authors cite data, I will first consider secondary publications. Do secondary authors cite or otherwise acknowledge the data they use? Table 4.4 lists the results of the classification process described above. By far, the most common method of acknowledging data produced by someone else is via a formal citation to that individual's (or group's) prior work on that data. Most of the papers in this sample (357 papers, 80%) included a citation of this type. Much less frequently (7 items,

2%), authors credited data producers informally, through an acknowledgement in a footnote or in the article text.

Table 4.4 Frequency of citation types across secondary publications (*N* = 449)

Type of citation	Number of papers
Formal data producer citation	357
Informal data producer citation	7
Formal data provider citation	99
Informal data provider citation	28

Citations to the data provider were uncommon. Under a third of the sample (127 articles, 28%) identified any kind of data provider. The majority of those that did provide a citation, though, included a formal citation (99 articles, 78%).

Table 4.5 Frequency of combinations of citation types across secondary publications (*N* = 449)

Combination of citations	Number of papers
Data producer citation only	278
Data producer plus data provider	86
Data provider citation only	41
Neither	44

Most papers cited the data producer and only the data producer (278 papers, 62%). Second most common, though, was to provide a citation both to the data producer and to the provider (86 papers, 19%). Notably, it was slightly more common to provide no citation at all than to cite only the data provider. (In these cases, authors just described the data they used, sometimes referring to it by title, but did not include the name of the producer or any indication of where they obtained the data.)

Citation norms change over time, and I found evidence of this in this sample. Over the time period represented by these papers, data citation has become more frequent. Secondary papers published after 2005 in this sample are significantly more likely to provide a data provider citation (formal or informal) than papers published previously ($X^2(1, N = 236) = 38.25, p < 0.001$).

Table 4.6 Proportion of papers citing data provider over time

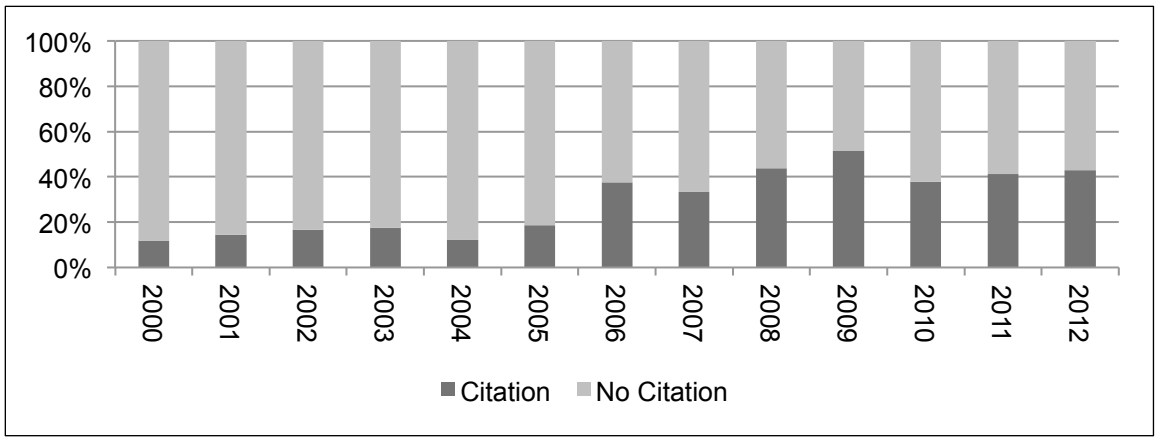
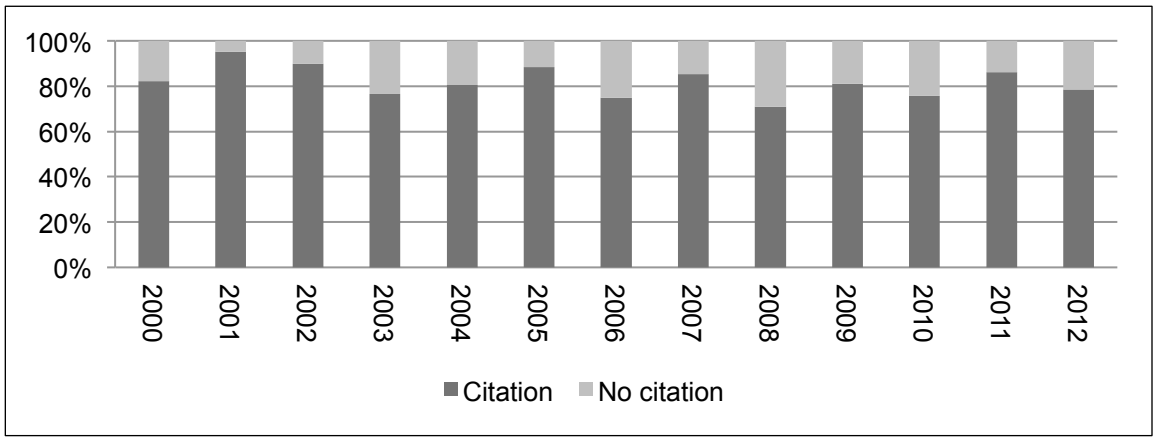


Table 4.7 Proportion of papers citing data producer over time



There is no evidence, however, of a similar pattern with citations to data producers. Later secondary publications are no more or less likely to cite the data producer than earlier publications.

4.4.2.2 Literature reviews, scholarly commentary and other data-related publications

The second part of this analysis focuses on the set of publications that are data-related but where the authors do not engage directly with the data—in other words, the papers are not analytical, or the analysis they contain is not centered on data.

Of the 176 items in these three categories, just 6 (3%, 2 literature reviews, 1 scholarly commentary, and 3 other data-related documents) provided a data provider citation of any kind. A slight majority, though, cited the data producer: 98 (56%) provided a formal or informal reference to the producer.

4.5 Discussion: Citation patterns in social science literature

Data producers do not usually cite the data that they produce or indicate whether it is publicly available or will be in the future. This is an interesting finding, in light of the fact that social scientists often discover new data for their work through reading the literature (Faniel, Kriesberg, & Yakel, 2012). The lack of formal citation practices among primary authors may reflect an underlying assumption that, unless otherwise specified, all data is available somewhere—if not in a repository, then through the data producer directly. A reader only need track down the data if she wants to use it. Alternatively, data producers often express a preference for personal contact with scientists who reuse their data (Birnholtz & Bietz, 2003); they may use a repository like ICPSR as an easy means of storage and dissemination, but prefer that reusers directly contact them to find out where to access the data.

By contrast, it is unusual for secondary users to fail to provide a citation of some kind, either to the data producer, the provider, or both. Most commonly, secondary users cite the data producer's earlier work, rather than citing the repository that provides access to the data. Mooney (2011) similarly found that data reusers habitually refer to PI's papers rather than to the repository in which the data are housed. She suggests that this is problematic, since reusers are not clearly acknowledging their scholarly debts. Citing the data producer's paper obfuscates whether the reuser is using data created by that person or just crediting her ideas or methods. However, this common finding—that data reusers frequently cite the producers rather than where they acquired the data—could be interpreted differently in light of Peritz's (1983a) citation classification scheme. Scholars use citations for multiple purposes, among them to acknowledge use of materials from other scholars. The fact that authors cite data producers' papers rather than repository sources may indicate that they perceive the 'material' they use when reusing data to be primarily embedded in the data producer's description of their research methods, etc. The data are, in a sense, ancillary to the process that created them, and as a result, reusers may preferentially acknowledge the data producer's work in producing data rather than work done by the data producer and the repository to share the data.

This practice, though, poses a challenge for tracking data reuse citations in the literature. A citation to a particular paper by the data producer could be an indication that the secondary user is using data, but it could also be just a regular background citation. While it is possible to make a somewhat-educated guess about what function the citation serves

based on where it is in the paper, the lack of distinction between citations-for-data and citations-for-literature makes automated retrieval of reuse citations difficult.

These findings suggest an emerging solution, though: the increasing uptake of formal data provider citations. Increasingly, authors are including citations to data providers (especially ICPSR, in this case) along with citations to data producers, or very occasionally, in place of them. Though this analysis was not designed to explore the impact of data citation mandates or guidance on data citation practices, the results offer some evidence that such efforts make a difference. I found that there was a significant increase in the proportion of papers providing a formal data citation after 2005, which may be related to a difference in how ICPSR presented its guidance on citing data. ICPSR had been providing data citations to data reusers since 1989, but in 2003, that information was made much more prominent on each study's information page (M. Vardigan, personal communication, May 10, 2013). In 2005, data citation text became a standard part of data downloads. The timing of these changes and the corresponding increase in data citations suggest that making citations visible and easy to use can encourage data reusers to include them in their papers.

Problematically, the earliest incarnations of ICPSR's data citations did not solve the automation problem, since they did not contain unique text. The citation format contained the data producer's name, the study name, the location it was produced, and information about ICPSR. The study ID number is unique within ICPSR's namespace (i.e., there is just one Study #6693 at ICPSR) but those 4-digit strings can appear coincidentally in

other contexts, especially URLs. There was no truly unique string that appeared consistently in every citation. The study information all appeared in other contexts frequently. Keying in on the data producer identification works in some instances, but not consistently: some users, for example, include the abbreviation “ICPSR” while others spell out the full name (sometimes with misspellings).

These issues point to the importance of unique identifiers for data. While providing information about who produced the data and where it came from is important, it is also critical to attach a unique key to data to enable citation tracking. ICPSR now assigns DOIs to individual studies, which provides a string that is uniquely tied to the data. The inclusion of DOIs in data citations, combined with the increasing number of authors who are aware of the need to provide formal data citations, will make bibliometric analysis of data reuse a more accessible possibility in the future.

Equally important to being able to track citations to data in the secondary literature is the ability to distinguish between instances of reuse and instances where authors are referring to data rather than using it. This analysis shows that authors already make that distinction. While data citation is increasing in secondary literature, and some primary authors cite data, in non-research articles like literature reviews, scholarly commentaries, and other data-related documents, authors rarely cite data providers, relying instead on citations to data producers.

These findings suggest that social scientists distinguish between different contexts in which data is used. Data producers do not, generally, cite their data: they have no need to be assigned credit for the data because they already get credit by describing and publishing their data in their papers. Noting that the data are or will be publicly available is, seemingly, not a priority. In particular, social scientists make a distinction between using data and talking about data. Secondary users, who draw on data in their analyses, regularly cite data in one way or another, giving credit most frequently to the original producer, but they increasingly often acknowledge the data provider as well. Authors of literature reviews, scholarly commentary or other materials who may be drawing on the ideas or results from someone else's study for background or to critique generally do not cite data. Instead, they cite the broader context of the data, especially the papers written by the data producers.

This chapter lays important groundwork for the analysis in the next chapter. In this chapter, I identified papers in which the data in my sample were reused and established that this sample of datasets adhered to previously observed citation patterns. This produced a corpus of citation data, which forms the basis of the analyses in the next chapter. The next chapter explores moves from identifying reuse to exploring patterns of reuse in more detail, particularly focusing on how to measure the impact of reuse through citation analysis.

CHAPTER 5

Measuring Impact

In this chapter, I move from identifying data reuse to quantifying the scholarly impact of that reuse. In the previous chapter, I drew together a set of publications that reuse my sample of datasets. Now, I introduce several different methods for counting the contribution represented by that reuse. I use two main approaches here: first, I developed a set of citation-based metrics in the vein of traditional bibliometric impact metrics. Second, I used an alternative, non-citation-based metric as a point of comparison. I then calculated all four metrics on my sample of datasets, producing four different rankings of the datasets from high to low impact. Finally, I compared the four sets of rankings to determine what value each metric had in providing a different conceptualization of data reuse impact.

5.1 Computing Impact Metrics from Citation Databases

Impact metrics, including those I introduce below, generally rely on tracking citations in published literature. While each of the metrics I developed here incorporates information about citations to data differently, at their core they rely on counting citations from reuse publications. This raises a crucial question: which citations should count? The ICPSR Bibliography of Data-Related Literature contains citations from books, book sections, reports, journal articles, theses, conference proceedings, audiovisual materials, and other documents. As reported in Chapter 4, the majority (68%) of the items in the bibliography

(after excluding duplicates and inaccessible items) are journal articles. The second largest group is reports, with 15% of the total. Books and book sections comprise only 7% of the citations to data. Should all of these citations be counted, or only some subset of them?

More often than in the natural and physical sciences, social scientists publish in a variety of venues and in that way are more similar in their publication practices to the humanities. A study of the output of the Spanish Scientific Research Council found that its humanities / social science division published 54% of its work in journals, compared to 81% in the natural science divisions (Pestana, Gomez, Fernandez, Zulueta, & Mendez, 1995). Australian social scientists publish journal articles or published conference proceedings about 61% of the time; natural scientists do the same about 85% of the time (Bourke, Butler, Biglia, & Australian National University. Research Evaluation and Policy Project, 1996). The remainder of the publications the social scientists in Bourke et al.'s sample produced comprises books, edited books, book chapters, monographs, reports, and other documents.

There are four distinct literatures in the social sciences, according to Hicks (2005): journal articles, books, national literatures (which she defines as journals and other venues geared toward a local community of scholars, especially those in languages other than English), and non-scholarly works. As a result, bibliometric analyses that focus solely on journal articles may miss other important aspects of social science research, such as the very high impact that books can have, or the impact of knowledge in application, which occurs in non-scholarly literature. One approach to measuring impact

in the social sciences would be to include citations to all types of literature (Nederhof, 2006). A complicating factor, though, is that books and journal articles form almost two distinct 'worlds' of citations. While books have historically been an important venue for the social sciences, receiving by some estimates as much as 40% of the citations in the social sciences (Broadus, 1971), citations within books tend to preferentially be to other books and non-journal article literature, and conversely, journal articles tend to preferentially cite other journal articles (Line, 1979). Since citation patterns differ significantly between books and journal articles, citations to data from journals and citations to data from books might not be directly comparable.

Further, the frequency with which researchers publish in books and journals is not uniform across disciplines within the social sciences. Psychology and psychiatry, along with other social sciences related to medicine and health, are much more similar to the sciences in their publication practices, with the bulk of their publication occurring through journals articles ("ISI Coverage by Discipline," 2005, p. 125). In more 'science-like' social science environments, traditional metrics relying on counting journal article citations will be more appropriate than they will be in more 'humanities-like' disciplines. There is some evidence that publication patterns are changing with time (and perhaps in response to pressures arising from increased usage of impact metrics that favor journal publication). A recent study found a pronounced shift in the proportion of publications appearing in journal articles versus books from what was previously observed. In the period 2000-2009, Engels et al. (2012) found that social science publishing was dominated by journal articles. Over 90% of the items published in each of 5 social

science disciplines (Criminology, Educational Sciences, Psychology, and Sociology) were journal articles (2012, p. 388). In only one area, Political Science, was the proportion of book publications more than 20% (2012, p. 388).

These findings are particularly pertinent for my study, since the sample of datasets on which I am focusing skews heavily toward 'science-like' social science disciplines, especially areas related to the health sciences. These findings also support to my decision to compute impact metrics solely based on citations from journal articles. Journals are an important publication avenue across the social sciences, increasingly more so in many disciplines, and especially in the disciplines represented among my sample of datasets. The abundance of journal articles and scarcity of books in the ICPSR Bibliography further suggests that journals are a key venue for data reuse publications, either because data reuse is more often published as an article or because disciplines that reuse data more frequently are those that preferentially publish in journals.

The final reason for including only journal publications relates to a different aspect of publication practices. In the social sciences, researchers often present work at conferences before publishing it (Harzing, 2013). Sometimes that preliminary work is published as conference proceedings, which does not preclude later publication of a revised version of same work as a journal article.[10] Journal articles also are frequently published from dissertations. And the reports in the ICPSR Bibliography are produced by

[10] Often, the later journal article will acknowledge earlier publication of proceedings on the same topic. See the first footnote in (Reisig & Parks, 2004) for an example of this practice.

researchers funded by the National Institute of Justice or other entities that require or encourage publication of a report of results as a condition of funding; these reports also often precede publication of a formal journal article. As a result, within the ICPSR Bibliography, these three publication types—reports, theses, and conference proceedings—often duplicate work published as journal articles. To avoid counting the same instance of data reuse multiple times, I am excluding these publication types.

A second major methodological consideration, having decided which citations to count, is to determine an appropriate source for additional citation information. The citations in the ICPSR Bibliography, and specifically the 449 reuse publications I identified in Chapter 4, serve as seeds for building citation networks stemming from those publications. The metrics below rely not only on counts of citations from the reuse publications, but from additional generations of publications as well (i.e., papers that cite reuse publications).

There are several major databases that index research papers and provide researchers access not only to the content of the articles, but to the linkages between them. In this dissertation, I use Scopus, an Elsevier product that indexes 19,500 journals in the sciences, social sciences, and humanities, the bulk of which are from 1996 and after (Elsevier B. V., 2012). In addition to interface displays of bibliometric information like h-index, journal impact, etc., Scopus also provides API access for automated retrieval of citation information for analysis, including author and affiliation information. I conducted a comparison of several major citation databases and determined that Scopus and Google

Scholar have the best coverage of social science literature in general and my sample of publications specifically; I chose to work primarily with Scopus due to its superior research tools and functionality. See Appendix B for an in-depth comparison of the different citation database options.

5.2 Measuring Data Reuse Impact

In this chapter, I introduce four metrics for measuring data reuse impact. The first three are citation-based metrics: reuse count, secondary impact, and diversity. The citation-based metrics address three different aspects of impact, respectively: (1) How much is the data reused?; (2) How important is that reuse?; and (3) How broad is reuse of the data?

Adapting 'citedness' (a count of how many journal articles cite a given paper) for datasets, I propose a measure I call 'reuse count.' Reuse count differs from citedness in that it specifically counts only citations from secondary publications; the number of publications produced by the original data producers does not directly affect this metric. Secondary impact looks beyond the number of reuse publications relating to a dataset to account for the impact those reuse publications have. This measure is essentially the g-index applied to datasets rather than to scholars or articles. The third citation metric, diversity, examines the distribution of reuse publications across subject areas. Finally, I calculate an alternative, non-citation-based metric, downloaders, which measures how many unique individuals access a dataset through ICPSR. Each metric and its derivation is summarized in Table 5.2. Full explanations, derivations, and analysis follow in the sections below.

Table 5.1 Summary of measures of data reuse impact

Item	Definition	Measure	Source
Reuse count	The total number of citations a dataset receives over a specified time period	Citation count: number of publications produced by authors other than the study team (2 years, 5 years, 10 years)	ICPSR Bibliography
Secondary impact	Scholarly impact of reuse publications	g-index: the number and citedness of reuse publications	Reuse publication information: ICPSR Bibliography Citation information: Scopus, Google Scholar
Diversity	'Spread' of data reuse into other disciplines	Rao-Stirling diversity: the distribution of reuse publications across subject categories External publication proportion: number of publications appearing outside dataset SC	Journal titles: ICPSR Bibliography Subject categories: Scopus
Downloaders	The total number of unique downloaders of a study over a specified time period	Count of unique downloaders of data (with or without documentation) (2 years, 5 years, 10 years)	ICPSR data download statistics

Though there are 273 datasets in my sample, only a minority (44 datasets, or 16%) have been reused. Reuse is the critical precursor for reuse impact, so the metrics in this section are only calculated for those 44 datasets.

5.2.1 Reuse count

The reuse count of a dataset is the number of reuse publications that cite it that are not written by the data producers. This metric differs from citedness in that it counts only citations from papers that reuse the data. In other words, this is the total number of

citations a dataset receives minus the number of citations that come from data producers and the number of citations from papers that do not engage in reuse, such as literature reviews. Reuse count is the most 'traditional' of metrics used in this study. This measure credits datasets with the amount of direct scholarly usage they receive: how many times were people able to produce new works of scholarship from a dataset? This conceptualization of impact aligns with standard ideas of value in scholarship; datasets that are high in reuse count are valuable in the same way that highly cited papers are.

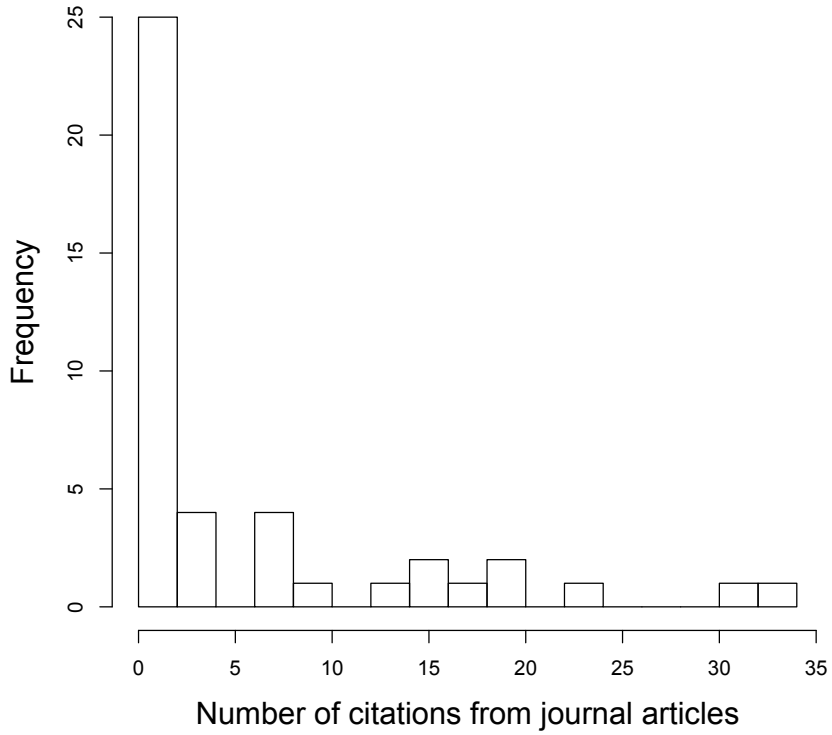
I base the reuse count of a dataset on the processing of the ICPSR Bibliography of Data-Related Literature described in Chapter 4. In that chapter, I divided the total set of data-related publications into primary papers, secondary papers, and other kinds of documents (including literature reviews, scholarly commentary and other miscellaneous documents). The reuse count for a dataset is the number of secondary papers related to that study. To compute this value, I created a Python script to produce a count of secondary publications for each study in my sample.

The script generated a year-by-year count of secondary publications based on the elapsed time between a paper's publication and the year in which ICPSR released the study that the paper is related to. I grouped the citations into three categories: pre-release publications (publications that were published before ICPSR's release of the data, or elapsed time since publication < 0); same-year publications (publication published in the same year the data were released, or elapsed time since publication = 0); and post-

release publications (papers published after the data became publicly available, or elapsed time since publication > 0).

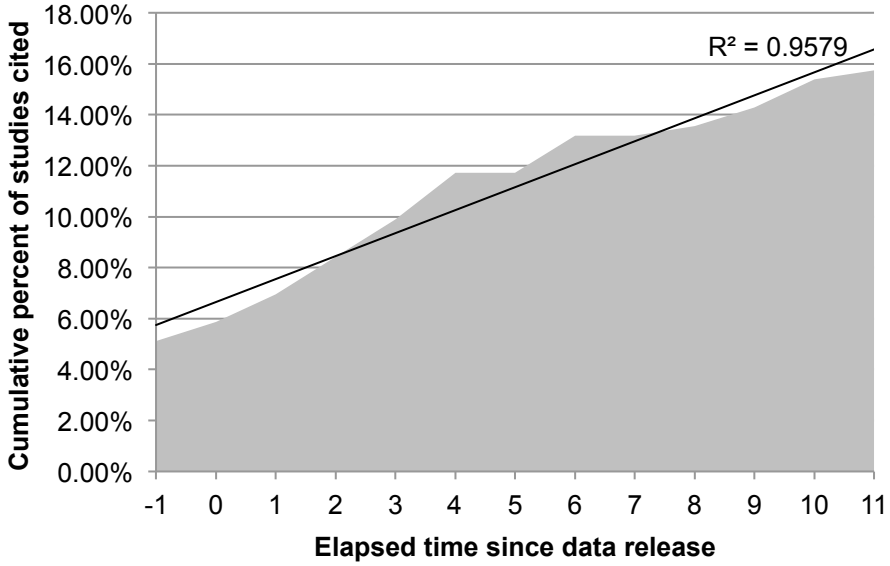
Of the total 273 studies in my sample, just 44 have reuse publications in journals. Figure 5.2 shows the distribution of the number of reuse citations. The majority of the studies, 229 (or 84%), have no secondary papers associated with them. Within the 44 studies that have one or more secondary citations, the number of citations ranges from 1 to 175, with a median of 1. The National Comorbidity Study (ICPSR #6693) is an extreme outlier, with 175 reuse citations. Of the 44 studies with reuse citations from journal articles, 33 (77%) have fewer than 10 reuse citations each.

Figure 5.1 Histogram of number of reuse citations from journal articles for studies with one or more citations (*N* = 44)



I also identified the first year in which a study received a reuse citation. Figure 5.3 shows the cumulative probability of reuse citation over 10 years (the period for which I have citation information for all studies in my sample). By the time 10 years have passed since ICPSR released the data, 15% of studies have received at least one reuse citation.

Figure 5.2 Cumulative percentage of studies cited (*N* = 273; 44 total studies cited)



Slightly under a third of the reused studies (14 studies, 33%) received reuse citations before the studies were released by ICPSR. Among those that were reused only after their ICPSR release, about three studies received their first reuse citation per year. The number of datasets cited increases linearly with time ($r(42) = 0.98$, $p < 0.001$). The strong relationship between time since release and number of datasets cited is an interesting finding. Data that are not reused within a few years of becoming publicly available are not necessarily doomed to obscurity. About a quarter of the datasets in this sample

(25.58%) received no reuse citations until 6 or more years had passed since ICPSR put them online.

Once a dataset is reused, it continues to be reused over an extended period of time. Of the 25 datasets with more than one reuse citation, just one received all its citations in a single year; ten of the 25 had reuse citations in five or more years.

The single most highly reused study in my sample is the National Comorbidity Study (ICPSR #6693). Table 5.3 lists the 10 highest impact studies by the reuse count measure:

Table 5.2 Top 10 highest impact studies according to reuse count

Study ID	Study Name	Reuse Count
6693	National Comorbidity Survey: Baseline (NCS-1), 1990-1992	175
3160	Project on Policing Neighborhoods in Indianapolis, Indiana, and St. Petersburg, Florida, 1996-1997	34
2884	National Treatment Improvement Evaluation Study (NTIES), 1992-1997	32
2851	Hispanic Established Populations for the Epidemiologic Studies of the Elderly, 1993-1994: [Arizona, California, Colorado, New Mexico, and Texas]	24
2258	Drug Abuse Treatment Outcome Study (DATOS), 1991-1994: [United States]	20
2976	Police Stress and Domestic Violence in Police Families in Baltimore, Maryland, 1997-1999	19
3385	Hispanic Established Populations for Epidemiologic Studies of the Elderly, Wave II, 1995-1996: [Arizona, California, Colorado, New Mexico, and Texas]	17
2833	National Survey of Adolescents in the United States, 1995	15
3337	Evaluation of the Gang Resistance Education and Training (GREAT) Program in the United States, 1995-1999	15
2778	Gambling Impact and Behavior Study, 1997-1999: [United States]	14

The distribution of the reuse citation measure is highly skewed: a few studies have most of the citations. The top three studies together comprise half the total number of citations (241 citations, or 54% of the 449 total citations in the sample[11]).

5.2.2 Secondary impact

Reuse count indicates how much a dataset is reused, but it says nothing about the quality of that reuse. How do scholarly audiences receive reuse publications? Are a dataset's reuse publications cited, or do they get ignored? A dataset may be cited infrequently and thus be low-impact by reuse count, but if the papers it produces are very high impact, the dataset should not be overlooked. The secondary impact measure makes it possible to distinguish a dataset that has a given number of low-impact reuse publications from a dataset that has the same number of reuse publications, but reuse publications of higher impact.

The secondary impact metric quantifies the degree of impact of reuse of a dataset via the g-index of its reuse publications. The g-index measures the impact of a set of publications where *g* is the number of highly cited articles within that set such that each highly-cited article has on average *g* citations. While the g-index and h-index are generally strongly correlated (Costas & Bordons, 2008, p. 283), a major strength of the g-index relative to the h-index in this context is that the g-index is not limited by the total number of publications in the body of work being considered. The h-index of a dataset is *n* if the

[11] Note that in this case, the total number of citations is equal to the number of publications. This is because each publication cites one dataset from the sample. But the papers could, in theory, cite more than one dataset from the sample, in which case the total number of citations would be larger than the number of publications.

dataset has *n* reuse publications garnering *n* or more citations. Consider two hypothetical datasets. Dataset A has 5 reuse publications that receive 5 citations each. Thus, Dataset A's h-index is 5. Dataset B also has 5 reuse publications, each of which receives 25 citations. Dataset B's h-index is 5, the same as Dataset A, though it could reasonably argued that Dataset B is the higher impact dataset because its total citation count is five times that of Dataset A. Because the h-index does not take into account the actual number of citations a paper has above the threshold *n*, it is not possible to distinguish the relative impact of these two datasets using that measure.

The g-index, by contrast, is not limited by the number of publications. Dataset A's g-index is 5, while Dataset B's g-index is 11. (See Appendix C for more detail on this calculation.) Costas and Bordons (2008) describe the g-index's particular appropriateness for measuring the impact of 'selective scientists,' researchers who publish relatively infrequently but receive high numbers of citations to the papers they produce. 'Big producers,' scientists who publish many papers (or in this context, datasets which have a large number of reuse publications), are favored by the h-index (2008, p. 283). Since this study already incorporates a measure, reuse count, that will highlight datasets with high levels of reuse, the g-index is especially appropriate here. If there are datasets that have few actual instances of reuse, but those instances are high impact, this measure will identify them.

The range of values for secondary impact is tied to the number of citations a dataset's reuse publications receive. As such, this metric does not have an upper bound. Though no

dataset will have less than zero secondary impact because it is not possible to have reuse publications that are cited fewer than zero times, reuse publications could be cited any number of times, and the secondary impact metric will increase along with the number of citations to reuse publications.

To compute the secondary impact measure for this group of datasets, I retrieved the total number of citations for each of the 449 reuse publications identified in Chapter 4. Most of these publications (378 items, 84%) are indexed in Scopus. For 66 articles, I pulled citation counts from Google Scholar. As noted above, Google Scholar consistently reported higher numbers of citations than Scopus. This discrepancy introduces a complication to the calculation of the secondary impact metric. Since Google Scholar's citation numbers were higher than Scopus's, the 66 articles for which I drew citation information from Google Scholar had inflated citation numbers relative to those with citation information from Scopus.

To estimate the amount by which Google Scholar's citation numbers were inflated relative to Scopus's, I returned to my test sample of articles, which I had looked up in all four citation databases. There were 41 articles in my sample that appeared in both Google Scholar and Scopus. Comparing the citation counts for these documents from Google Scholar directly with those from Scopus, Google Scholar consistently reported higher citation counts. In five cases, the citation counts reported by both databases were equal. In the other 35 cases, Google Scholar reported citation counts between 8% and 233% higher (*M*: 69% higher; median: 60% higher).

To make the numbers I got from Google Scholar comparable to the numbers from Scopus, I needed to weight the Google Scholar values. I used a weight based on the relationship between Google Scholar's citation counts and those provided by Scopus, which I determined through linear regression. The two sets of counts are strongly related (Adjusted $R^2 = 0.96$, $F(1, 39) = 864.8$, $p < 0.001$). The regression equation has an intercept of 0.60 and a coefficient of 0.62; I used this equation to produce weighted estimates for the citation counts of the 66 articles missing from Scopus.

Four articles (0.89%) were not in either citation database, and these articles are excluded from the analysis in this section. Three of those articles were related to datasets with more than one reuse publication, while one was the sole reuse publication for its dataset, ICPSR #6540 (White-Collar Criminal Careers, 1976-1978: Federal Judicial Districts). Because I could not determine a citation count for this article, I could not calculate secondary impact for this dataset.

Table 5.4 displays the top 10 datasets by secondary impact. The National Comorbidity Survey again appears at the top of the list, with a secondary impact of 83. The median secondary impact within these 43 datasets is 6.

Table 5.3 Top 10 highest impact datasets by secondary impact

Study ID	Study Name	Secondary Impact
6693	National Comorbidity Survey: Baseline (NCS-1), 1990-1992	83
2258	Drug Abuse Treatment Outcome Study (DATOS), 1991-1994: [United States]	21
2851	Hispanic Established Populations for the Epidemiologic Studies of the Elderly, 1993-1994: [Arizona, California, Colorado, New Mexico, and Texas]	20
2884	National Treatment Improvement Evaluation Study (NTIES), 1992-1997	20
2778	Gambling Impact and Behavior Study, 1997-1999: [United States]	18
3385	Hispanic Established Populations for Epidemiologic Studies of the Elderly, Wave II, 1995-1996: [Arizona, California, Colorado, New Mexico, and Texas]	17
3450	Pennsylvania Sentencing Data, 1998	17
3160	Project on Policing Neighborhoods in Indianapolis, Indiana, and St. Petersburg, Florida, 1996-1997	16
3337	Evaluation of the Gang Resistance Education and Training (GREAT) Program in the United States, 1995-1999	15
2833	National Survey of Adolescents in the United States, 1995	14

Three datasets (ICPSR #3212, #3240, and #3482) have zero secondary impact, reflecting the fact that while these datasets were cited in reuse publications, those reuse publications themselves were never cited.

5.2.3 Diversity

One of the promises of data preservation and reuse is that data collected to serve one purpose could be used by scholars to serve another, even in an entirely new area. The diversity measure identifies datasets that are used broadly. Are there different communities of researchers publishing from the data, or is reuse limited to a single field?

Previous work on research diversity has been aimed at measuring the amount of interdisciplinarity within research output. Porter et al. (2007) demonstrate that the network analysis techniques can reveal the degree to which an individual researcher's body of work is interdisciplinary. ISI Web of Science (WoS) classifies journals into different Subject Categories (SC); every journal indexed by WoS is associated with at least 1 and up to 6 SCs. To determine the interdisciplinarity of a researcher's papers, Porter et al. used a measure called Rao-Stirling, which involves computing the number of SCs represented among the body of work as well as the relatedness of that set of SCs. Researchers who placed papers in a greater number of SCs and/or in SCs that were less related to one another were determined to be more interdisciplinary. The same technique has been used to determine research diversity at the level of an organization (Soos & Kampis, 2010) or an entire field of study (Porter & Rafols, 2009).

Betweenness has also been proposed as a bibliometric measure of interdisciplinarity (Leydesdorff, 2009), as well as the Gini coefficient and Shannon entropy, both well-established measures of diversity from other fields. However, comparisons of all these metrics found that no measure stood out as the single best method for characterizing interdisciplinarity (Kajikawa & Mori, 2009; Leydesdorff & Rafols, 2011). Furthermore, these measures of interdisciplinarity are difficult to extend to datasets because datasets would hold a unique place in a citation network. For example, the betweenness of a dataset within a network of reuse publications is not a useful metric of comparison. A network of a single generation of reuse publications stemming from a dataset is a star graph, and so the dataset will be on all shortest paths. This means that the betweenness

centrality value for the dataset will be directly correlated with the size of the network and adds no meaningful information about the diversity of that network.

My goal was to quantify the diversity among the set of publications that directly reuse data, i.e., the diversity of a dataset's reuse publications. Therefore I selected Rao-Stirling diversity as the most effective measure in this context. Rao-Stirling diversity balances three elements: variety, balance, and disparity (Stirling, 2007). Variety is the number of types within a sample, balance is the distribution of elements across the types, and disparity is a measure of how different the types are from one another.

The formula for Rao-Stirling diversity is

$$\delta = \sum p_i p_j d_{ij} \qquad \text{(Eq. 1)}$$

where p_i is the proportion of elements in category i; p_j is the proportion of elements in category j; and d_{ij} is the distance between the two categories. The diversity is the sum of this product over all categories. The range of this metric is determined primarily by the values used for the distance between categories, d_{ij} and the number of categories the collection of elements represents: the metric will increase both with an increase in the number of categories and the distance between them.

In this case, the categories I used are subject categories (SCs) as assigned by Scopus' Scimago Journal and Country Rankings. I assigned SCs for the reuse publications such that for each publication, the SC is the SC or combination of SCs of the journal in which the article was published. Table 5.5 shows the number of reuse publications per SC.

Table 5.4 Reuse publications (*N* = 449) per subject category

Subject Category	Num. reuse pubs.
Medicine	116
Medicine + Psychology	74
Social Sciences	59
Medicine + Social Sciences	48
Psychology + Social Sciences	38
Psychology	18
Medicine + Pharmacology, Toxicology and Pharmaceutics	17
Medicine + Psychology + Social Sciences	11
Medicine + Neuroscience	11
Medicine + Nursing	8
Medicine + Nursing + Psychology	8
Business, Management and Accounting + Medicine + Psychology + Social Sciences	5
Medicine + Pharmacology, Toxicology and Pharmaceutics + Psychology	5
Mathematics + Psychology + Social Sciences	3
Biochemistry, Genetics and Molecular Biology	2
Economics, Econometrics, and Finance	2
Economics, Econometrics, and Finance + Social Sciences	2
Business, Management and Accounting + Social Sciences	2
Medicine + Neuroscience + Psychology	2
Medicine + Pharmacology, Toxicology and Pharmaceutics + Social Sciences	2
Ag. and Bio. Sciences + Economics, Econometrics, and Finance + Soc. Sciences	2
Neuroscience	1
Nursing	1
Business, Management and Accounting + Economics, Econometrics, and Finance + Social Sciences	1
Medicine + Nursing + Psychology + Social Sciences	1
Medicine + Nursing + Social Sciences	1
Biochemistry, Genetics and Molecular Biology + Medicine	1
Medicine + Neuroscience + Pharmacology, Toxicology and Pharmaceutics	1
Biochemistry, Genetics and Molecular Biology + Medicine + Psychology	1
Computer Science + Medicine	1
Environmental Science + Social Sciences	1
Computer Science + Social Sciences	1
Decision Sciences + Mathematics	1
Biochemistry, Genetics and Molecular Biology + Medicine + Psychology + Social Sciences	1
Interdisciplinary	1

There were 449 reuse publications, 315 (70%) of which belong to journals whose SC's are Medicine (alone or in combination with other fields). This compared to 215 (48%) with Social Sciences (alone or in combination) classifications. Just 4 articles (1%) had journals with no Social Science or Medicine SCs.

To complete the Rao-Stirling diversity calculation, I needed to derive the distance between each pair of SCs in the sample. Leydesdorff and Rafols (2011) found that Rao-Stirling diversity is sensitive to how distance is calculated; however, they also found that there was no best option among the methods they explored. Here I calculate distance within a network of SC cocitation data from Scopus, using the unweighted shortest path length between each pair of SCs. The Scopus data provided cocitation between individual SCs but did not include the interdisciplinary categories I describe above as nodes (i.e., Social Sciences and Medicine are both nodes, but Social Sciences + Medicine is not). To determine the distance between interdisciplinary SCs, I computed the shortest path as above for each component of the interdisciplinary SC and then averaged the values. For example, the distance between Computer Science and Psychology is 4, and the distance between Computer Science and Medicine is 3, so the distance between Computer Science and Medicine + Psychology is 3.5.

I created a Python script to compute Rao-Stirling diversity, Eq. 1 above. The results for the 10 highest diversity studies are listed in Table 5.6.

Table 5.5 Top 10 highest impact datasets by Rao-Stirling diversity

Study ID	Study Name	Rao-Stirling Diversity
3334	Aging, Status, and Sense of Control (ASOC), 1995, 1998, 2001 [United States]	0.91
3023	Cooperative Agreement for AIDS Community-Based Outreach/Intervention Research Program, 1992-1998: [United States]	0.75
2884	National Treatment Improvement Evaluation Study (NTIES), 1992-1997	0.69
3136	Continuity and Change in Criminal Offending by California Youth Authority Parolees Released 1965-1984	0.64
2258	Drug Abuse Treatment Outcome Study (DATOS), 1991-1994: [United States]	0.64
3337	Evaluation of the Gang Resistance Education and Training (GREAT) Program in the United States, 1995-1999	0.62
2778	Gambling Impact and Behavior Study, 1997-1999: [United States]	0.61
2835	National Pregnancy and Health Survey: Drug Use Among Women Delivering Live Births, 1992	0.56
3190	National Organizations Survey (NOS), 1996-1997	0.54
3355	Recidivism of Prisoners Released in 1994	0.52

Rao-Stirling diversity is high when a dataset's reuse publications are distributed across multiple subject areas and low when those publications are concentrated in one or a few areas. However, there is an additional dimension of diversity that this metric does not take into account. There were 9 datasets with zero diversity. All 9 of these datasets had all of their reuse publications in a single SC. However, for 7 of those 9 datasets, the SC in which their publications appeared was different than the dataset's SC. Since the metric was designed to measure diversity within a collection of publications, it does not account for the data's own SC and the distance between that subject area and the areas in which reuse publications appear. Thus, a dataset with all reuse publications in a single SC will have zero diversity, regardless of whether that SC is the same or different than the SC of the dataset itself.

Rao-Stirling diversity rewards datasets when they are used to produce publications in multiple subject areas; however, a dataset can be interdisciplinary if used in only one subject area if that subject area is different than the dataset's subject area. To accommodate this type of interdisciplinarity, I propose an extension to the Rao-Stirling diversity formula:

$$\delta = \sum p_i p_j d_{ij} + \sum p_k d_k \qquad \text{(Eq. 2)}$$

The first component remains the same as in Eq. 1. The second component adds the contribution of the proportion of publications p_k in SCs at some distance d_k from the dataset's SC. This diversity metric rewards datasets for a broad distribution of reuse publications among SCs as well as for publications outside the dataset's SC. Table 5.7 lists the top 10 datasets by this combined diversity measure.

Table 5.6 Top 10 highest impact datasets by diversity (adapted Rao-Stirling)

DataID	Study Title	Diversity
3190	National Organizations Survey (NOS), 1996-1997	2.85
3334	Aging, Status, and Sense of Control (ASOC), 1995, 1998, 2001 [United States]	2.59
2884	National Treatment Improvement Evaluation Study (NTIES), 1992-1997	2.14
3023	Cooperative Agreement for AIDS Community-Based Outreach/Intervention Research Program, 1992-1998: [United States]	1.95
2851	Hispanic Established Populations for the Epidemiologic Studies of the Elderly, 1993-1994: [Arizona, California, Colorado, New Mexico, and Texas]	1.92
3337	Evaluation of the Gang Resistance Education and Training (GREAT) Program in the United States, 1995-1999	1.80
3385	Hispanic Established Populations for Epidemiologic Studies of the Elderly, Wave II, 1995-1996: [Arizona, California, Colorado, New Mexico, and Texas]	1.79

3355	Recidivism of Prisoners Released in 1994	1.74
3163	Harvard School of Public Health College Alcohol Study, 1997	1.71
2854	State Court Organization, 1998: [United States]	1.67

The new values for the diversity metric are quite different than those produced by the Rao-Stirling formula. As expected, the diversity values for each dataset are equal to or higher than the Rao-Stirling values; since the new formula adds a term to the Rao-Stirling formula, it will never produce lower values. All but 5 datasets had at least one publication outside of their own SCs, which indicates that the disparity between a dataset's SCs and those of its publications is an important dimension to capture.

There was no relationship between the number of reuse publications a dataset had and its diversity score ($r(41) = 0.198$, $p = 0.20$). There were datasets with little reuse but relatively high diversity, and conversely, datasets with many reuse publications but low diversity. ICPSR #3334 (Aging, Status, and Sense of Control (ASOC), 1995, 1998, 2001 [United States]), for example, has just four reuse publications, but each of those publications is in a different SC (Nursing, Social Sciences, Medicine + Nursing, Medicine + Neuroscience + Psychology). On the other end of the spectrum, ICPSR #6693 (National Comorbidity Survey: Baseline (NCS-1), 1990-1992) has 175 reuse publications, but close to 70% (124 publications) were in just two closely related SCs (Medicine and Medicine + Psychology), one of which was the same as the dataset's own SC (Medicine + Psychology).

Factors that might influence diversity include the amount of time data have been available and the dataset's SC. I hypothesized that older data or data from an interdisciplinary SC would have higher diversity. The longer data have been available, the more opportunities may have arisen for researchers from various fields to discover and use the data; similarly, more researchers may be exposed to data that is part of an interdisciplinary SC. However, neither hypothesis was supported within this sample. The diversity metric is not normally distributed ($W = 0.95$, $p = 0.03$), so I used a Kruskal-Wallis test to assess the relationship between the year the data were released and the diversity metric. There is no relationship between the year the data were released and the diversity metric (Kruskal-Wallis $X^2(26, N = 44) = 28.20$; $p = 0.35$). This does not rule out a relationship between time and diversity, but suggests that the two years between the release of the oldest and newest data in this sample is not sufficient to show a difference in diversity after 10 years.

To test whether data from multi- or interdisciplinary subject areas have higher diversity, I divided the subject area assignments into two groups: single-discipline and multidisciplinary. There was no relationship between whether data were from a single discipline or multiple disciplines and their diversity scores (Kruskal-Wallis $X^2(26, N = 44) = 31.01$; $p = 0.23$).

5.2.4 Downloaders

The metrics above are all based on citations, but there are other bases for examining how and how effectively a piece of scholarship is used. Some journals are beginning to offer

alternative and article-level measures of impact, generally referred to as 'altmetrics.'[12] One alternative metric some publications are beginning to offer is the number of downloads a paper receives. Though download counts do not appear to correlate with citation counts for Open Access articles (Davis, Lewenstein, Simon, Booth, & Connolly, 2008), download statistics are useful in showing increased readership of those same articles compared to closed access articles. Similarly, high numbers of downloads of a dataset may not indicate especially high levels of citation, but download counts may provide a useful lens into uses of the data that do not result in a citation. For example, a user may download a dataset to use as training data for an instrument or analysis, without incorporating it into their main published analysis, and thus may not cite the dataset, or an instructor may have students download and work with data as part of an assignment. This kind of impact may be less visible than reuse of data that gets directly cited in a paper, but it may be just as important, particularly for repositories, researchers, and funders that are interested in education or in fostering reuse outside the academy. Datasets that are downloaded by many people but rarely cited, for example, will fall through the cracks of traditional metrics like reuse count, but they may have significant value for repositories to preserve.

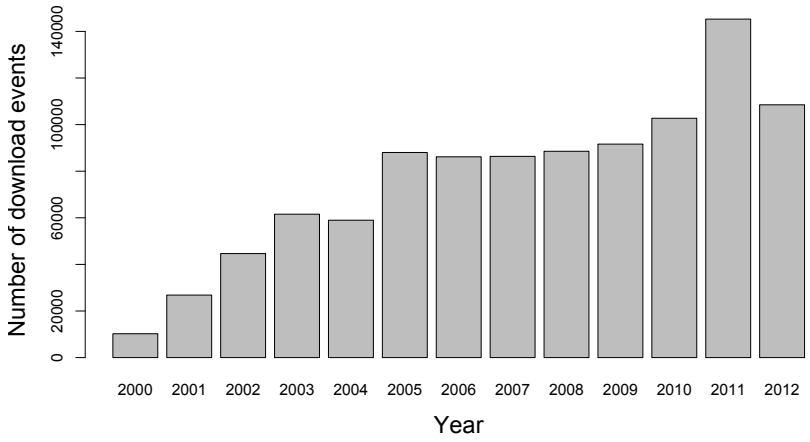
Here, I compute the number of downloaders of each dataset in my sample. This metric provides a contrast to the citation-based metrics of the previous sections. I extracted all download events from ICPSR's web statistics, which ICPSR stores in a SQL database. A

[12] See for example, PLOS (http://article-level-metrics.plos.org/alm-info/), Cambridge Journals Online (http://blog.journals.cambridge.org/2012/04/latest-cjo-development-article-level-metrics/), and BMJ Open (http://blogs.bmj.com/bmj-journals-development-blog/2011/05/27/article-usage-metrics-now-available-on-bmj-journals/)

download event is recorded every time a user downloads any part of a study. There were 1,173,873 download events related to the data in my study sample from Jan 1, 2000 (the year direct downloads became available) through Dec 31, 2012. Since there is not yet complete data for the year 2013, I excluded 2013 downloads from this study.

The number of download events has climbed each year since ICPSR started allowing direct downloads (Figure 5.4). There was a substantial increase in downloads in 2005. In that year, the download interface changed to make it easier for users to download all files relating to a study with a single click. This bulk download function delivered multiple versions of all data files (SAS, Stata, Excel, etc.), the codebook, and any other associated files at once. Many users took advantage of the opportunity to download everything all at one time rather than picking and choosing which files they wanted. Though they accomplished this with one action, the download statistics reflect downloads of each individual file.

Figure 5.3 Download events by year (*N* = 1,173,873)



Accordingly, not all of the 1,173,873 download events I identified are actually of data. For example, study #2258 (Drug Abuse Treatment Outcome Study (DATOS), 1991-1994: [United States]) has 90 separate files associated with it. There are 10 distinct datasets within the study, and most of these datasets have a codebook file, a questionnaire file, files with four different data types (SAS, SPSS, Stata and ASCII) plus three data setup files (e.g. Stata do files, etc.). A user downloading this entire study will increase its file download count by 10, whereas a study with just one data file will have a count increase of just one, even though in both cases, the downloader pulled the entire study. Furthermore, a user may download data files multiple times (to get a clean copy after making changes to an original download, for example). To avoid the influence of both duplicative downloads as well as differences arising from different numbers of data files across studies, in this study, I use counts of unique downloaders rather than file downloads.

To download data from ICPSR, users must log in to their 'MyData' account with a registered email address. Of the 1,173,873 download events, 745,413 (64%) were by registered users. A further 290,688 (25%) events were through a 'guest@icpsr.umich.edu' email address. In the course of their work, ICPSR staff may occasionally download data or other documentation; about 1% of the download events (14,707) occurred through emails with an 'icpsr.umich.edu' domain.

Other materials, like codebooks, are available for download without logging in. There were 76,982 anonymous download events (7%), and 46,083 (4%) download events with

only an IP address recorded. In 2000, 2001, and 2002, anonymous downloads made up the majority of download events (100%, 89% and 72%, respectively), dropping to 6% in 2003 and less than 0.01% thereafter.

Users may download datafiles only, datafiles and documentation, or just documentation. To further refine the measure, this study will exclude downloaders who only ever downloaded documentation files. These users never acquire the data from ICPSR. While it is possible that a portion of these users acquired the data from another source (direct contact with the data producer, for example, or from another repository), there is no way to know how large a proportion those users represent, so they will be eliminated from consideration here.

To identify unique downloaders, I excluded anonymous downloads, downloads through the 'guest@icpsr.umich.edu' login, and ICPSR staff email addresses, which left only download events through registered emails and through recorded IP addresses. Then, I de-duplicated the sets of registered email addresses and IP addresses. The final result comprises 791,496 download events by 88,693 unique downloaders.

Because anonymous downloads (which were excluded) are the majority in the years 2000, 2001, and 2002, it is possible that datasets released earlier are penalized in my calculation of this metric. In this sample, datasets in 2000 are essentially missing three years of downloader data while datasets released in 2002 are only missing one year. If most or a large proportion of downloaders access data in the first few years after a dataset is released, the metric would systematically undercount the number of downloaders for

datasets released in 2000 and 2001 compared to those released in 2002. According to a Shapiro-Wilk test, the downloaders metric is not normally distributed ($W = 0.58$; $p < 0.001$), so I used a Kruskal-Wallis test to determine whether there is a significant difference in the downloader metric for studies released in 2000, 2001 and 2002. This test showed no significant effect of release year on the download metric (Kruskal-Wallis $X^2(209, N = 273) = 207.66$, $p = 0.51$).

Table 5.8 shows the top 10 highest impact studies according to the downloaders measure. The downloaders metric ranges from a maximum of 3,787 (National Comorbidity Study, #6693) to a minimum of 0 (four studies: Harlem Longitudinal Study of Urban Black Youth, 1968 United States, ICPSR #121; National Congregations Study (NCS) United States, ICPSR #122; CrimeMapTutorial Workbooks and Sample Data for ArcView and MapInfo, 2000, ICPSR #3143; and Regional Crime Analysis Geographic Information System (RCAGIS), ICPSR #3372). The median is 188.00.

Table 5.7 Top 10 highest impact datasets by downloaders

Study ID	Study Name	Down-loaders
6693	National Comorbidity Survey: Baseline (NCS-1), 1990-1992	3,787
2790	World Values Surveys and European Values Surveys, 1981-1984, 1990-1993, and 1995-1997	3,393
2778	Gambling Impact and Behavior Study, 1997-1999: [United States]	2,637
3088	Alcohol and Drug Services Study (ADSS), 1996-1999: [United States]	2,478
3355	Recidivism of Prisoners Released in 1994	2,209
2833	National Survey of Adolescents in the United States, 1995	1,984
3334	Aging, Status, and Sense of Control (ASOC), 1995, 1998, 2001 [United States]	1,658
3226	Homicides in New York City, 1797-1999 [And Various Historical Comparison Sites]	1,482

2859	Height of Runaway Apprentices and Military Deserters in Colonial and Early Republican America, 1726-1825	1,291
2835	National Pregnancy and Health Survey: Drug Use Among Women Delivering Live Births, 1992	1,252

The overall distribution of downloaders is in Figure 5.5.

Figure 5.4 Distribution of downloaders metrics

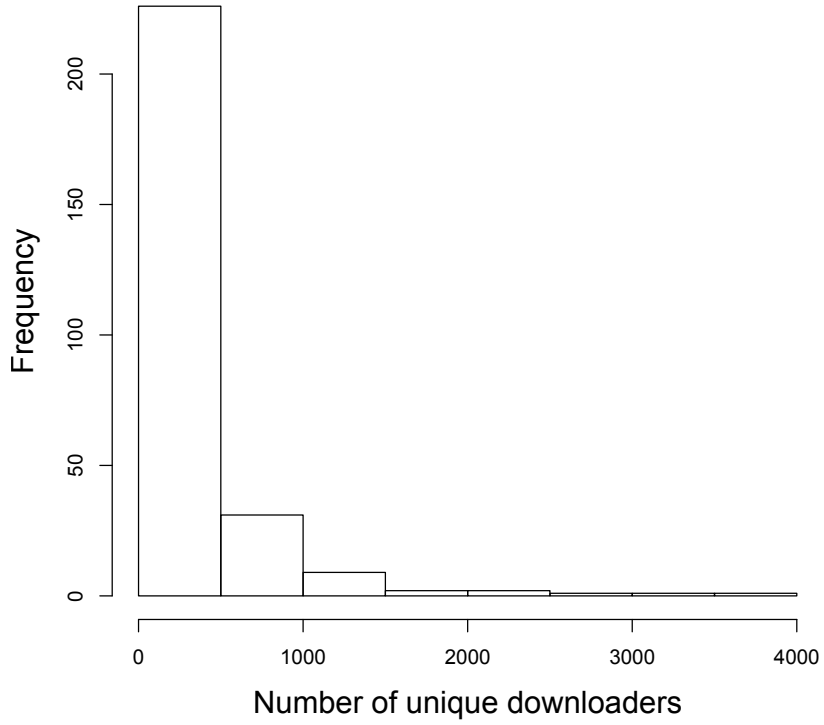


Figure 5.6 shows the relationship between the time elapsed since the data were released by ICPSR and the proportion of new unique downloaders gained per year. Each cohort follows a distinct pattern: an increase in the number of unique downloaders gained per year to a rate that is maintained for about 5 years, then an additional increase. But the cohorts are time-shifted from one another. As Figure 5.7 shows, the changes in the rate of new downloaders gained seem to correlate more with the calendar year than with the amount of time the data have been available.

Figure 5.5 Median new unique downloaders by year post release

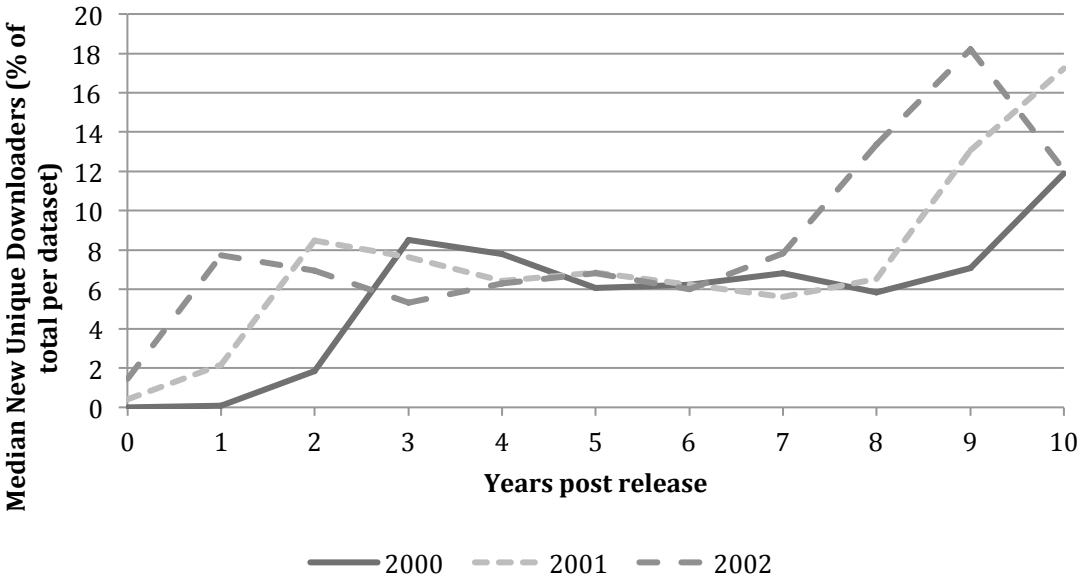
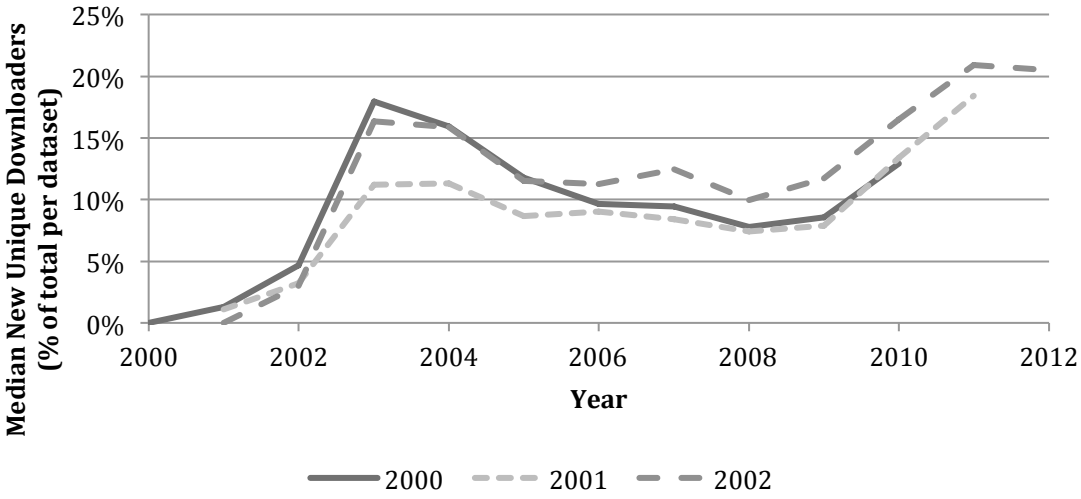


Figure 5.6 Median new unique downloaders by calendar year



5.3 Discussion: Comparing Metrics

One goal of this research was to compare the usefulness of these several measures of impact. Do the same datasets receive high impact measures using each and every

method? Or do these different measures of impact highlight different datasets? Are the most downloaded datasets also the most cited datasets, and the ones that produce the most highly cited papers and so on? Or are there instead datasets that are very frequently downloaded but rarely cited? Or infrequently cited, but the papers are highly regarded?

There are 44 datasets for which I was able to generate a value for all four metrics. The limiting factor is reuse citations: of the 273 datasets in my sample, only 44 datasets had reuse publications. Appendix D lists all 44 datasets and their rankings according to each measure.

Because there are ties within each set of rankings (i.e., two or more datasets with the same score on a measure), I used Kendall's tau-b to compare the sets pair-wise. Table 5.9 shows the correlation coefficient for each pair.

Table 5.8 Correlations between data reuse impact measures

	Reuse Count	Secondary Impact	Diversity	Downloaders
Reuse Count	—	0.957	0.174	0.653
Secondary Impact		—	0.223	0.649
Diversity			—	0.342
Downloaders				—

There are significant correlations between reuse count, secondary impact and downloaders, with reuse count and secondary impact being the most closely related.

Diversity, however, was not significantly related to any other metric, suggesting that this metric captures a different type of impact than the other three measures.

There are data that are high-impact almost across the board. An example of this is the ICPSR #6693, the National Comorbidity Study (NCS). This study is the most reused in the sample, as well as the data with the highest secondary impact, and the most downloaded. Further, it is in the top third of datasets by the diversity measure. In this case, multiple metrics do not add much information. The NCS was reused 175 times in journal articles, in addition to the 83 papers published by the data authors. The sheer amount of work stemming from this data is evidence that it represents an important scholarly contribution.

Few datasets scored so highly by the reuse count metric; many more scored very low. One interpretation of this result is that these datasets, in contrast to the NCS, are not especially valuable for reuse and thus do not have much of an impact beyond their original usage. An alternative interpretation, though, is that the raw number of instances of reuse is not the best yardstick for data impact. While the reuse count metric is easy to calculate, it does not provide much discriminatory power. Using additional metrics makes it possible to examine the impact of these datasets more granularly, and specifically, to untangle low frequency of reuse from low impact. A key finding in this chapter is that reuse count is not an ideal metric for this sample of datasets, and to the extent that these data are representative of data held at ICPSR and social science data more generally, reuse count may not be an especially useful metric for such data.

Metrics that are not based solely on frequency of reuse shed light data that is not reused much but has impact in other ways. Table 5.10 lists three studies that score highly (in the top quartile) by one metric but lower on the other three.

Table 5.9 Outlier datasets

		Ranking			
Study ID	**Study Title**	*Diversity*	*Down-loaders*	*Secondary Impact*	*Reuse Count*
2835	National Pregnancy and Health Survey: Drug Use Among Women Delivering Live Births, 1992	30	**7**	23	20
3450	Pennsylvania Sentencing Data, 1998	30	29	**6**	14
3190	National Organizations Survey (NOS), 1996-1997	**1**	14	18	20

Study #2835, the National Pregnancy and Health Survey: Drug Use Among Women Delivering Live Births, 1992, demonstrates the largest discrepancies of the three highlighted studies. The study ranks seventh by the downloaders metric, but barely above the 50th percentile on any other metric. This suggests that this is a study that is used by many people, but that usage is not reflected in the published journal literature. These data may be used in a classroom setting, for example, or by journalists or policy makers who do not write journal articles from their work with the data. It is clear that these data have value to users, but that value is not embodied in the kinds of scholarly impact identified in the other three metrics.

Similarly, Study #3190, the National Organizations Survey (NOS), scores highly on diversity but relatively low otherwise. Conversely, Study #2778 (Gambling Impact and Behavior Study, 1997-1999) scores low on diversity but high on all other metrics. This illustrates a nuance of the diversity metric. While secondary impact and reuse count clearly identify 'high' and 'low' impact datasets, those categories are not quite correct for the diversity measure. Instead, the diversity measure identifies 'broad' or 'narrow' reuse; while 'high' and 'low' imply a clear value judgment, the same is not true for 'broad' versus 'narrow.' Datasets with especially broad impact are interdisciplinary and valuable in multiple areas, but datasets that are especially narrow may be particularly important to the subject in which they are reused.

These examples illustrate the main value of developing multiple impact metrics. These metrics address different aspects of Costas et al.'s (2013) conceptual landscape of data metrics and, in doing so, ameliorate some common problems with metrics. Aspects of impact that are not measured can be unfairly minimized, and conversely, those that are easily measured can be artificially inflated. Together, this suite of metrics provides a more nuanced view of impact: not only can one identify data that many people use, but also data that are valuable to one field or many, or even beyond academic purposes as well.

This kind of broad view on impact is important given that metrics are often used in decision-making: author impact measures are frequently used in hiring and promotion considerations, and similarly, data impact measures may in the future be used in the same

contexts. Furthermore, repositories could use impact metrics to assess the strengths of the data they curate and to set priorities for ongoing curation and preservation.

This chapter addressed the important issue of how to quantify impact and introduced several possible metrics for doing so. Furthermore, I used these metrics on a sample of social science datasets to demonstrate their effectiveness in identifying high-impact data. An important consideration for repository managers in particular, though, is not only to be able to look retrospectively at the impact of their collections, but to be forward looking as well. The next chapter uses the metrics developed in this chapter as outcome variables and turns to the question of what factors influence how datasets score via different measures of impact. How can we tell which datasets will be reused? What makes reused data different than the rest?

CHAPTER 6

Anticipating Impact

The previous chapters identified data reuse in the social science literature, and used that citation information to develop impact metrics for data. Measuring impact retrospectively is important; however, repositories, funders, and researchers have a vested interest in understanding how to anticipate impact prospectively. In this chapter, I move toward the goal of understanding predictors of data impact in three steps. First, I review the literature on scientists' decisions around reusing data to identify a set of potential factors that influence reuse. Then, through an analysis of interviews with actual social science data reusers, I refine the set of factors to those that are most likely to be relevant within a social science context. Finally, using a series of regression models, I identify the factors that correlate with increased odds of reuse and higher reuse impact with my sample of ICPSR datasets.

6.1 Anticipating impact by anticipating data reuse

While there have been no systematic studies of predictors of data reuse, there has been some study of a similar problem: what makes a researcher likely to be successful in the future? When it comes to evaluating researchers, tenure committees make educated guesses about who will be strong contributors in the future by looking at their previous publication records, with whom they have worked, the amount of funding they have

secured, and other factors. Acuna, Allesina and Kording (2012) were the first to systematically study which of these characteristics are actually predictive of future success. They collected a large dataset of information about a group of neuroscience researchers, such as the number of articles published in top journals, the number of years in a postdoctoral position, average number of coauthors per paper, and others, for a total of 18 features. Using a linear regression, they developed a model that identified five key features that predicted the researchers' future h-indexes: h-index, total articles published, years since first article, number of distinct journals published in, and number of articles in 'top' journals.

A similar approach could be used to predict the impact a dataset might have. There are features of authors that predict how high an impact the papers they produce will have; are there features of datasets that predict how high impact the papers produced from them will be? Are there certain characteristics of datasets with high impact that are distinct from datasets that are not? Data impact fundamentally arises from data reuse; if a dataset is never reused, it has no reuse impact, regardless of whether impact is measured as a direct count of reuse or whether it assesses downstream repercussions of reuse as described above. Factors that make a dataset likely to have high impact, then, should be a subset of the factors that make a dataset likely to be reused. AddHealth is an example of a study that produces an exceptionally high level of reuse and thus has high impact. What is it about this study that makes it possible for so many people to reuse it in so many ways? Is it the sheer size of the dataset (9,258 variables)? The fact that it is well

documented and easily accessible? Is it that the data are unique? Does it have to do with the data producers who happen to be established and respected researchers in their field?

In this dissertation, I follow Acuna et al.'s model and use regression analysis to explore correlates of data reuse rates. Specifically, I developed a set of factors that may facilitate data reuse and tested them in models using datasets' reuse outcomes and scores on the impact metrics I introduced in Chapter 5 as the outcome variables.

6.2 Influences on data reuse

Numerous studies have examined barriers to and facilitators of data reuse, and these studies suggest factors that may affect which datasets are heavily reused and which are not. These studies do not all agree; they do not all find that the same factors have an influence on whether data are reused or not. While there is some consensus, no single factor or set of factors emerges as a crucial barrier or facilitator to reuse. In developing the set of characteristics that will serve as the independent variable in my model, I took a broad view of the literature, including factors as they appeared and generating hypotheses based on the conclusions of the cited studies below.

Niu (2009) found that data users prioritize their information needs: if a dataset is what they need to do their research, they will find a way to use it, whether it is documented well or not. And conversely, even if a dataset is beautifully documented, if it is not what a user needs, they will not use it. This finding indicates the importance of considering characteristics specific to the dataset: what it is and what it contains are keys to whether it gets reused. Larger and more comprehensive datasets are likely to be at an advantage

here: a dataset that is large enough to be carved up in numerous ways is more likely to suit a variety of research questions than one that has fewer variables, a smaller *n*, or a tighter focus on a single topic.

Multiple studies highlight the importance of information about how data were collected. Zimmerman (2008) finds that the ability to visualize the data collection process is key to ecologists' data reuse; similarly, Faniel and Jacobsen (2010) found that earthquake scientists pay particular attention to descriptions of the experimental design of a study when determining whether that study's data are acceptable for their own reuse. Based on these findings, datasets with documentation that specifically details the data collection process will be more likely to be reused than those with documentation that either does not include that information or that presents it unclearly.

Niu (2009) found that quantitative data, and especially surveys, tend to be better documented than other kinds of data. Carlson and Anderson's case study of several research groups found that the group engaged in survey research documented its data collection processes in greater detail than researchers in the other disciplines that they studied (2007, p. 65). They attributed this to the relatively large size of the survey research group and the distribution of tasks across the members of the research team. Because different individuals were responsible for different aspects of the data collection, it was crucial for all members of the team to be explicit about what they had each done and how they did it. That documentation could be included with the data when they were archived, which in turn aided potential reusers in understanding the data. The number of

data producers a dataset has, then, may influence the likelihood that those data are reused by necessitating clearer and more complete documentation.

Another set of elements that may influence whether a dataset is reused and how—and by extension, the impact that dataset has—has to do with the data producer. Potential reusers bring with them knowledge both of their field and its norms as well as knowledge about the people in that field. Their personal knowledge of the producer (or the producer's reputation) of data they are considering using may influence whether they choose to use the data or not. Van House et al. (1998) found that the reputation of a data producer could positively affect data reuse. Zimmerman (2008), however, found that producer reputation was not an important factor, except in cases where potential reusers had a very low tolerance for uncertainty in the data, or when there were questions about the data collection process; in those cases, knowledge of the skill level of the data producer came into play, though not as a deciding factor.

Van House et al. (1998) further suggested that reuse is more likely to occur when the data producer and a potential reuser are part of the same discipline. Some of the work that goes into data collection and processing is subject to the tacit knowledge problem: knowledge a member of a particular community gains over time about how to do things in a generally accepted matter is often not encoded in the data or documentation directly. Knowing that the data producer is part of a shared community of practice allows the reuser to take it for granted that the data producer followed the set of unwritten rules by which the community abides. As Van House notes, when the producer is not part of the

community of practice, reusers may be less able to assume that the producer followed the accepted protocols (1998, p. 341).

Birnholtz and Bietz (2003) and Edwards et al. (2011) suggest another way information about the data producer can influence reuse. Both studies identified informal, ad hoc communication and social connection between a producer and a reuser as an important part of successful data reuse. As Edwards and his colleagues explain, no matter how structured and thorough metadata about a dataset is, metadata products necessarily lack total precision, and as a result, friction arises in the sharing and reuse processes. Data reusers and producers engage in conversation and discussion to eliminate that friction. Social connections are important, and so pre-existing connections between a potential reuser and a data producer may positively influence whether a dataset gets reused. The importance of pre-existing connections in turn suggests that the size of data producers' social networks may play a role in whether the data they produce is reused. If a data producer has previously worked with a large number of people, it is more likely that potential reusers have either worked directly with the producer before or are only a few degrees separated which in turn may make them more likely to reuse the data.

The actions of repositories and the changes a repository enacts to data can also make a difference in whether or not the data are reused. Daniels et al. (2012) developed a taxonomy of changes based on interviews with repositories that serve three different disciplinary communities: social science, archaeology and zoology. The repositories studied change data to add value to them, to correct errors, to create consistency, to make

data reflect new knowledge, to respond to the specific needs of the repositories' designated communities, and to reflect changes in the way that data are collected. These actions are intended to make it easier for reusers to access and use data. Data that is held by a repository that engages in such actions, then, may more be likely to be reused than data that is either not part of a repository or that is held by a repository that only distributes data exactly as it was submitted.

Table 6.1 Summary of literature on data reuse

Study	Field studied	Findings
Zimmerman (2008)	Ecology	Ability to understand local context of data collection (to 'see through' to the data collection activities) is critical to reuse Personal knowledge of data collector does not influence decision to reuse, though knowledge of that individual's skill level can make a difference
Faniel and Jacobsen (2010)	Earthquake science	Potential reusers scrutinize experimental design of a study to evaluate its data for reuse.
Niu (2009)	Social science	Data users prioritize information needs over documentation quality; if data are relevant to their research question, they will use the data, but if the data are not a fit, they will not.
Van House et al. (1998)	Environmental planning	Data users establish trust in data via knowledge of the collecting organization's reputation, their knowledge of the individual data producer (personally or by reputation), or by knowledge that the producer shares membership in a disciplinary community
Edwards et al. (2011)	Climatology	Where metadata products lack precision, they induce friction in data sharing. To overcome this friction, scientists engage in metadata processes, or in other words, use ad hoc conversations and other measures to reduce friction.
Birnholtz and Bietz (2003)	Earthquake engineering,	Data are crucial at the boundaries of communities; they mark the edges of

	HIV/AIDS research, and space physics	communities of practice, but can also be used to gain entrée into a field and to mark one's status in it. Sharing supplemental materials and supporting social interaction around data will promote sharing and reuse.
Carlson and Anderson (2007)	Astronomy, survey research, museum collection development, anthropology	Sharing data is an act of communication to an unknown audience, which requires making the objects of communication (data) transportable and intelligible. Quantitative data are particularly well-equipped for abstraction and transmission. This is even more true for data produced by large project teams; when many hands are involved in collecting data, there is a greater need for explicit descriptions of what each person did and how, and this documentation helps reusers to reconstruct those processes.

With a few exceptions (Niu, 2009; Daniels et al., 2012), the studies cited above focus primarily on the physical and life sciences. Social scientists do their work differently than physical or life scientists; what bearing might findings in these fields have on reuse in the social sciences? First, the life and physical sciences are quite different from each other as well. Lab-based research and field-based research are arguably as different from each other as the social sciences are from either, but these studies show some commonalities across the different fields. This suggests that there may be some overlap into the social sciences as well. Furthermore, there is not yet evidence that these findings are specifically *not* relevant to the social sciences. If the factors identified in other fields turn out in this study to have no bearing on how social scientists reuse data, this would be an important finding.

6.3 Influences on data reuse in the social sciences

From the literature described above, I identified five primary factors that may influence scholar's reuse of data, summarized in Table 6.2. Niu (2009), who studied quantitative social scientists, emphasized research fit. This is a challenging aspect of data to measure, since research fit will be different according every scholar's needs. But a proxy for research fit might be the size of the dataset (Factor A): larger datasets (with more variables or more cases) may be suited to a greater variety of information needs than a more focused or smaller dataset. Another factor is whether there is information available to users about how the data were collected (Factor B). This information helps potential reusers evaluate the data in a number of ways: determining whether the data are a fit for their research questions (Niu, 2009), whether they can get a complete enough understanding of the data to work with them (Carlson & Anderson, 2007), and evaluating to what extent the data are trustworthy or have been collected according to acceptable standards (Zimmerman, 2008; Faniel & Jacobsen, 2010).

Potential reusers may also rely, albeit in a more limited fashion, on information about data producers when deciding whether data are trustworthy enough to be useful to them. Data producers with better reputations in their fields (Factor C) may produce more reliable data (Van House et al., 1998; Zimmerman, 2008); similarly, data producers who are part of a disciplinary community (Factor D) may be expected to adhere to the norms of that community when producing their data (Van House et al., 1998). Finally, reuse may be more likely when data reusers are able to connect directly with data producers:

personal contact and communication with data producers facilitates the process of data reuse (Birnholtz and Bietz, 2003; Edwards et al., 2011).

Table 6.2 Factors that influence data reuse

Factor	Justification
A. Dataset size (Niu 2009)	Datasets that are large may be more likely to suit the information needs of a greater number of potential reusers, and thus may be more likely to be reused.
B. Data collection process information (Niu 2009; Zimmerman 2008; Faniel and Jacobsen 2010; Carlson and Anderson 2007)	Datasets with more information about how they were collected may be more understandable to users, especially novice users, and so may be more likely to be reused.
C. Data producer reputation (Van House et al. 1998; Zimmerman 2008)	Data producers with a better reputation may be perceived to create more reliable or better data.
D. Discipline of data producer (Van House et al., 1998)	A shared disciplinary membership between the data producer and reuser allows the reuser insight into the producer's data practices.
E. Connection with data producer (Birnholtz and Bietz, 2003; Edwards et al. 2011)	Personal connections between producers and reusers foster reuse because reusers can ask questions about the data as well as build trust in the data and their producer.

6.3.1 Interviews with social science data reusers

In order to assess whether the factors described above are relevant in a social science context, I conducted a secondary analysis of interviews with social science data reusers originally collected as part of Elizabeth Yakel's and Ixchel Faniel's Dissemination Information Packages for Information Reuse (DIPIR) project (Institute for Museum and Library Services grant # LG-06-10-0140-10).[13] The DIPIR project is an Institute for

[13] www.dipir.org

Museum and Library Services (IMLS)-funded three-year project to explore data reuse practices in three different disciplinary communities: quantitative social scientists, archaeologists, and zoologists. The DIPIR investigators are approaching their research questions through a variety of methods: interviews, observation, focus groups, and survey research. I am repurposing their set of 43 interviews with social science data reusers. These interviews were structured around participant's descriptions of their experiences reusing data: how they decided what data to reuse, where they found them, how they determined whether the data were trustworthy, etc. The interviews also addressed participant's perceptions of data repositories and data trustworthiness in general.

The interviews were conducted between June 2011 and April 2012 either by phone (19) or in person (24). Both DIPIR investigators and three graduate assistants (myself included) conducted the interviews. Of the 43 interviewees, 40 had actually used data previously; three respondents expected to reuse data in their future work and answered the interview questions prospectively. About half the interviewees (22) were novices, primarily Ph.D. students, and half were experts, primarily professors or professional researchers. The novice interviewees were recruited through flyers and publicity during ICPSR's summer program. To recruit experts, we relied on referrals from ICPSR's Official Representatives, who serve as official coordinators of ICPSR's services at their institutions, snowball sampling of interviewees, and other professional contacts. All interviewees were compensated $25 for their time.

I used a qualitative data analysis software, NVivo10 (QSR International, 2012) to manage the text from the interviews and my coding scheme. I began coding by using a basic codeset comprised of the five factors described above. As I read through the transcripts, I used an inductive approach to identify additional themes (Miles & Huberman, 1994). After completing an initial round of coding, I reviewed my codeset and added new codes to correspond to the new themes that had emerged. As a last step, I recoded the interviews according to the finalized codeset.

The next sections describe my results. They are arranged by code, with the first five section corresponding to the initial five factors I derived from the literature, and the final section summarizing an important new theme that emerged from the interviews.

6.3.2 Research Question Fit / Dataset Size

The first set of features that may play a role in whether a dataset is reused or not is made up of characteristics of the dataset itself, specifically those relating to the size of the dataset. Datasets that are large may be more likely to suit the information needs of a greater number of potential reusers, and thus may be more likely to be reused (Niu, 2009).

Several reusers of ICPSR data highlighted the importance of research question fit. The DIPIR respondents typically began to look for data with a research question or model to test already in mind. CBU10 had considered what she would ask if she was doing data collection on her own and chose to work with a dataset that included similar questions. Others described having determined what their key independent and dependent variables

would be and used codebooks and other information to determine whether those variables were present in the data (CBU02, CBU24, CBU26, CBU39, CBU40). If the dataset did not have those key variables, respondents noted that sometimes they could adjust their research questions or analysis somewhat (CBU12, CBU15), but more likely would need to look elsewhere for usable data (CBU18). If the data had the right variables, a question still remained, though, as to whether there was sufficient data to conduct the desired analysis, especially when the research question depended on a specific subsample of the overall population (CBU26, CBU40, CBU43)

Though the size of a dataset may relate to its suitability for multiple research questions, reusers of ICPSR data did not frequently talk specifically about dataset size in relation to research question fit. When considering whether to reuse a particular dataset, these interviewees' primary concern about data size was logistical: large datasets can be unwieldy to work with (CBU15, CBU26). CBU12 noted that sometimes smaller is better, since finding a targeted dataset will save time one might otherwise spend searching through a very large dataset to identify the few variables actually needed. Respondents cared about the number of cases a dataset had because a larger number of cases can compensate to some extent for missing data. If there are missing values in a smaller dataset, removing those cases may not leave enough data to run an analysis. For CBU16, a better option would be to find a dataset that is not as good a fit for her question but has less missing data than to attempt to work with a dataset that has too much missing.

Larger datasets may promote reuse because they are more flexible and suit a wider range of research questions, but they may also inhibit reuse because they can be difficult to work with. In this study, dataset size will be captured in two ways: the number of variables in the study, and the number of cases. Both of these values come from ICPSR's documentation for the data; in the cases where the documentation did not contain that information, I downloaded the data and identified the number of variables and cases manually.

6.3.3 Information about the data collection process

DIPIR interviewees confirmed previous research findings that information about the data collection process was key to enabling their reuse (Niu, 2009; Zimmerman, 2008; Faniel and Jacobsen, 2010; Carlson and Anderson, 2007). It was particularly crucial for potential reusers to be able to discern the exact measurements and operationalizations used in the data, and in survey data, the question wording employed (CBU03, CBU10, CBU15, CBU16, CBU17, CBU18, CBU22, CBU33). CBU32 described a need to see the process "from the original data collection to the numbers that you've got in front of you"; CBU34 noted that if good process information is not in a dataset's codebook, "I get angry and stop using it."

Interviewees expressed pre-existing ideas about what to look for when examining the process underlying a given dataset. CBU23 has "really specific expectations about certain types of behaviors" that are acceptable when creating data, and for him, knowing the details of how the data came to be was as important as knowing the actual content of the data. CBU04 looked for adherence to sampling standard procedures; if the data producer

followed commonly accepted practices, CBU04 would feel that she could trust the data producers' measurements. For CBU10, it was important to compare how the data he wanted to use were actually collected to what he would have done if he had the chance to collect them himself.

Information about the data collection process was important for two closely related reasons. Data are specifically constructed to measure a quantity in relation to a research question. When a reuser tries to repurpose that data, explained CBU13, "[y]ou can easily measure something else than you think you're measuring if you don't really understand what [the data producer] decided" about how to structure the questions she asked and the survey overall. In a sense, effective reuse of data is not possible without in-depth knowledge of the original purpose of the data. For this reason, many of the respondents mentioned checking their understanding of the data's structure, variables, and processing by replicating known results (CBU05, CBU06, CBU29, CBU31). For CBU36, the process of replication also helped establish the trustworthiness of the documentation in and of itself.

In addition to gaining functional knowledge about how the data work and what they can be used for, information about the data collection process also helps reusers to discern whether the data are trustworthy or not. Scrutinizing how the data were collected can help reusers identify any skew or bias in the data (CBU10, CBU11). CBU09 pointed out that data producers' own biases can be embedded in the data, either intentionally or unintentionally: "I think everyone knows that when it comes to data, there's a lot of ways

to play around with it to make your point. And that I would want to make sure that I'm not using data that doesn't help me because it's incorporating those forms" (CBU09).

Studies that have been processed by ICPSR include some information about how the data were collected in the study metadata; studies that are part of the Publication-Related Archive were not processed and have no data collection process information included. The additional information recorded for processed studies could help potential reusers gain more insight into the data collection process, and so I hypothesized that processed studies are more likely to be reused than unprocessed studies, and furthermore, that this factor will be significant in the regression model of reuse.

6.3.4 Data Producer Reputation

Under certain conditions, especially when data reusers have strict needs for the quality of the data they will use (Zimmerman, 2008), the stature of the data producer can influence whether data will be reused. The DIPIR interviewees primarily discussed reputation in terms of the institution rather than the individual researcher. CBU12's comments reflected the overall sentiment of the respondents:

> I have a hierarchy where I try to get the information from governments if I can, state government and federal governments, sort of repositories. And then, if there are certain groups like the fiscal survey of the states, I don't think is a federal initiative, but it's a group... It's a... I don't know if it's not nonprofit, but it's a group that's been doing it for a long time, so it's well-recognized. And then, any data that I get from other political scientists, I mean, I have to trust it. [chuckle] It's good data based on their reputation. (CBU12)

Many respondents expressed their preference for data from government sources or from an institution with a good reputation (CBU09, CBU10, CBU11, CBU19, CBU25,

CBU29, CBU36, CBU39). CBU36 noted that the government data she works with are not documented especially well, but she is more tolerant of that with this government data than she would be otherwise. Since she trusts the government to produce good data, the documentation does not need to be quite as thorough.

When it comes to individuals, CBU37 relied "on the collective judgment" about someone's reputation; CBU22 similarly commented that as "you started working with survey data you built certain familiarity with who collects data and whether using their data is good or bad." CBU19 and CBU28 both turned to the published literature to get a sense of a data producer's reputation. If a data producer has been published in good journals, then she is more credible than someone who has not.

Data producer reputation is captured for non-institutional authors through their h-index. I parsed author names from ICPSR's DDI metadata and coded them for type as described in Chapter 3. Using Scopus's author registry, I retrieved each individual author's h-index at the time of data release. For groups of authors, I averaged their individual values to arrive at a single value for the group.

6.3.5 Connection with data producer

Birnholtz and Bietz (2009) and Edwards et al. (2011) suggest that personal connections with a data producer can facilitate reuse. In Birnholtz and Bietz's study, data producers expressed a preference for arranging reuse of their data through direct contact, rather than through an intermediary like a data repository. When reusers have contact with producers, it gives them a chance to ask questions about the data and to build trust with

the producer. However, only three interviewees mentioned ever asking questions of the producers of the data they used. One of those three, CBU35, noted that when the documentation is good, there is no need to talk to the data producer directly.

However, contact with data producers did affect data reuse in a different way: direct or indirect ties (through an advisor, for example) to a data producer are a good way to find out about useful data and to gain access to them (CBU05, CBU06, CBU11, CBU12, CBU16, CBU30, CBU31, CBU44). Personal connections are an important route for information transfer and sharing in particular in the social sciences (Line 1971), and for these interviewees, this dynamic holds true for information sharing about data as well.

These connections may play a role as well in how often data are reused and by whom. Potential reusers have a greater chance of being connected with a data producer if a study has many authors and/or authors who have large co-authorship networks. In this study, I use the size of a data producer's co-authorship network to capture their connectedness to other scholars.

6.3.6 Discipline of data producers

Another aspect of the potential reuser's connection with a data producer is whether they share a disciplinary affiliation (Van House et al., 1998). However, the DIPIR interviewees barely mentioned this factor. One (CBU11) addressed discipline, somewhat indirectly, by discussing an institution's reputation in a given field as a reason to trust or distrust data. Another noted that he wanted to branch out into another field and would use

data that established researchers in that field use in order to gain entrée into that research area.

These limited remarks suggest that, on the whole, the data producer's discipline may not be not a crucial factor in reusers' decisions about reuse. I identified the datasets' disciplines based on the disciplines of their data producers through the process described in Chapter 3. Since there are a large number of categories with relatively few datasets per category, I recoded the data to capture whether the data producer has a single disciplinary affiliation or multiple. Data with multidisciplinary producers may be more likely to be reused since the producers have connections in multiple communities; alternatively, producers with a single discipline may have more and deeper connections within that single community, which could foster increased reuse by a particular set of scholars.

6.3.7 An additional factor: data in the published literature

An important emphasis from the interviews and one not explicitly discussed in the literature on data reuse is social scientists' use of literature to identify, locate, and assess the quality of data that is available for reuse. As part of their everyday work, social scientists engage with the literature in their fields. Reading publications is an important avenue for social scientists to be exposed to new data that could be useful to them. CBU33 described how she came to use the data she does:

> I found about them because other scholars in my field were using those datasets, and I heard about them just from reading other people's work, and I realized that there might be useful information on the types of questions I wanted to ask too. (CBU33)

Simply by engaging in the process of doing social science research, a scholar gets to know what data is available for use and what data is worth using. As CBU25 stated, "People tell you what to use," and becoming familiar with data that are commonly (or even uncommonly) used is "part of the research process." Within research areas, certain data "is going to keep popping up" (CBU26).

Discovery of new data through the social science literature can happen actively or passively. While the interviewees quoted above describe the process of becoming exposed to new data somewhat passively—read lots of papers and you will get to know what data to use—for students in particular, finding new data can be an active task. Professors or advisors may suggest that a student identify existing data to use rather than try to collect their own. CBU10 was guided away from original data collection and given a number of articles to read by his advisor; that reading introduced him to the data that he now intends to use in his thesis. CBU19 describes the process of actively seeking out new data. Starting with a literature search on his topic in JSTOR, he noted, "I'd find authors or I'd find articles from the top journals, and saying, 'Okay, if it made American Political Science Review and American Journal of Political Science, it's probably a pretty good paper. What dataset did they use?'" Reviewing the literature related to data, either passively or actively, also allows social scientists to determine whether data will be useful for their purpose, both in terms of research fit and quality of the data. CBU19's response gets at the quality dimension of identifying data through the research literature. He was more interested in data that has been published in high-profile journals, using the quality of the paper as a proxy for the quality of the data.

The presence of publications related to data can indicate their quality in other ways as well. The existence of multiple publications related to data indicate that those data have been used a number of times, and importantly, that a number of people have checked the data and the original producers results. CBU41 notes that "through repeated use, you know, people notice things," that could be problematic in the data. If there are issues with data, the user is likely to contact the original producer, who would correct the issue. The more publications there are, the more use the data have received, and the more likely any errors have been ferreted out. CBU16 sets the threshold quite high, though: if they are cited "by more than a hundred people," the data can be trusted. For others, like CBU15 and CBU32, who rely on data that is less-used, the relatively few papers related to their data serve the same purpose. CBU33 explained in more depth that if data are frequently used in her field (criminology), she knew that she was "not going to get halfway into the dataset or analyses and then find out, oh, there was some major problem that [she] hadn't recognized." Additionally, knowing that data were generally accepted within the field removed some uncertainty about the methods used to collect the data. CBU33 knew that because many people have already published from the data she uses, she would not face "big fundamental questions" from reviewers about the data and the design of the instruments used to collect them.

But exposure is not the only function literature serves. It continues to be useful once a scientist has decided to use the data: published papers might act as an additional source of documentation about the data or a supplement to the codebook. For CBU24, publications

provide guidance on how to work with data. Though codebooks can provide variable descriptions and other important information, CBU24 looks at publications "to know if there's a standard way of [combining variables], be it in a composite or a scale for example, and that's the number one thing I look at. And then also a lot of articles will tell you things like, well, you can't do this or you shouldn't do this." For simple datasets, related publications may be less important, but for those that are "a little more complicated" (CBU05), they can provide a bit of a shortcut to busy researchers, "[b]ecause sometimes that's easier just to look at their method section and the variables they looked at than to sort through all the codebooks and everything yourselves" (CBU33).

The amount of literature—primary and secondary—about a dataset cuts across categories. In these publications, potential reusers find information about data collection process, potential research question fit, how the data were used, how the data cannot or should not be used, who created them, and who has used them. Acuna et al.'s (2012) study of predictors of researcher h-index found that the best predictor of an individual's future h-index is their current h-index. Success breeds success, in some ways. Data reuse may be subject to a similar pattern. The prominence of data in the research literature at a given point in time may influence how likely it is to be reused later on. It may be the case that high numbers of publications indicate that a dataset is especially valuable; the publications may also simply serve as publicity for the data. Either way, the likelihood of others using the data may increase.

6.3.8 Institutional vs. individual authors

This definition of factors has so far skimmed over a crucial point: most of the datasets in my sample were authored by a single or group of authors, and for many of those that were not, I was able to identify the individuals behind the organizational author. But 46 datasets have only an institution or government agency as their author, and for these authors, h-index and coauthor network size are meaningless. Further, there is a significant relationship ($X^2(1, N = 273)$, $p < 0.001$) between processing status and type of author: all institutional or government-authored studies are processed. So the processing status variable is not meaningful either in the case of institutional or governmental authors. This leaves primary publications before reuse as the only variable of interest for institutional or government authors. But this is problematic: government and institutional datasets tend toward administrative data or other data-gathering efforts that do not result in primary publications. In fact, just six of the 46 institutional or government-authored datasets had any primary publications at all. This suggests that these datasets are not suited to the analysis I planned here. The social scientists in the interviews analyzed above described trusting data from government or institutional sources based on their own knowledge of those entities; that type of reputation or, in a sense, social capital is not captured in any of these factors. For that reason, I limit the sample in the next section to the 227 datasets authored by individuals or groups of individuals.

6.4 Choosing regression models

Table 6.3 summarizes the initial set of independent variables to be used in this study:

Table 6.3 Initial set of independent variables: predictors of reuse

Independent variable	Measure	Hypothesis
A. Data collection process information	Processing status (0 or 1)	Studies that have been processed by ICPSR have information about the data collection process in their metadata, which will make those studies more likely to be reused.
B. Data producer reputation	h-index of data producers (individual or average)	Datasets produced by high-profile, high-expertise individuals or groups will be more likely to be reused.
C. Connection with data producer	Size of data producers' co-authorship networks	Potential reusers may be more likely to use a dataset if they have a preexisting working relationship with the producer. The size of data producers' co-authorship networks will positively correlate with reuse.
D. Prominence of data	1. Usage: Number of publications authored by data producers 2. Prior reuse: Number of secondary publications	Datasets that are popular, either because they are heavily published from by their producers or because they are reused frequently, may continue to be reused at a higher rate than other datasets.
E. Size of dataset	1. Number of variables 2. Number of observations	Broader datasets with more variables are more likely to be useful in answering a greater number of research questions than a small, narrowly focused one. Studies with more observations may similarly support more kind of analysis, because more observations enable more ways to subset the data. Larger datasets – more observations or more variables – will be more likely to be reused.
F. Discipline of dataset	Multidisciplinary (0 or 1)	Multidisciplinary studies may have higher levels of reuse because their producers are connected to multiple communities; alternatively, single discipline studies could have more reuse because their producers have more and stronger connections within one field.

There were two primary questions I aimed to model: what factors influence whether or not a dataset was reused, and what factors influence reuse impact.

For logistic regression, a sample size rule of thumb is a 10 cases per predictor, with a minimum of 100 total cases (Harrell, Lee, Califf, Pryor, & Rosati, 1984) though some have argued that a less stringent rule is sufficient (Vittinghoff & McCulloch, 2007). For the model of whether data are reused or not, there are four predictors and 227 cases, so the sample size should be sufficient, at least at first glance.

I conducted a power analysis for a multiple logistic regression using Stata's powerlog function. I calculated the effect size for the variable that I hypothesize to be most important: the number of primary publications. At the mean, the proportion of datasets that have been reused is 0.09; at one standard deviation above the mean, the proportion is 0.30. Table 6.4 shows the required sample size to achieve 80% power at a significance level of 0.05 and effect size of 0.21. The squared multiple correlation between the predictors is unknown, so I assessed the sample size requirements over a range of values.

Table 6.4 Sample size calculation at varying squared multiple correlation values, with power = 0.8, alpha = 0.05 and effect size = 0.21

R2	Sample size
0.1	167
0.3	215
0.5	301
0.7	501
0.9	1504

If the predictors are not well correlated, my sample size is sufficient, but at greater correlations, the sample size is small. Small sample sizes result in biased estimates, specifically a systematic bias toward higher regression coefficients and higher odds ratios. Logistic regression conducted on a small sample will tend to overstate the effect of the predictors (Nemes, Jonasson, Genell, & Steineck, 2009). Exact logistic regression will ameliorate this problem, but it is computationally intensive.

A further problem is that, within this sample, reuse is a rare outcome (31 of the 227 datasets were reused), which introduces additional problems (King & Zeng, 2001), among them complete separation. Complete separation occurs when a predictor or combination of predictors perfectly predict the outcome. King and Zeng suggest a method to address biased estimates in small samples which is similar to the more broadly used method of penalized likelihood, also known as Firth logistic regression (Heinze & Schemper, 2002). Firth regression addresses both the problems of small sample size and potential complete separation, so I use it here to model what factors predict whether or not a dataset is reused.

To assess the relationship between these factors and reuse impact, I created models with the impact metrics as the outcome variables. For reuse count, secondary impact, and diversity, the sample is comprised of only the datasets that have been reused. This is a small sample, so I recoded the outcome variables into high/low binary variables divided at the mean of each distribution and conducted an exact logistic regression.

For downloaders, the sample need not be restricted to reused datasets, since not all downloaders produce scholarly publications. While count data often follows a Poisson distribution, downloader count also displays signs of overdispersion: the variance (149,432.80) is far larger than the mean (298.24), suggesting that the data fit a negative binomial distribution. To model downloaders count, I used negative binomial regression, an ordinary count model.

Table 6.5 Descriptive statistics: Continuous variables

	Number of variables	Number of observations	h-index (release)	Co-authorship network size	Number of primary pubs before reuse
Mean	572.85	10314.78	4.46	45.59	1.89
Std. Dev.	1499.18	42543.86	8.51	75.77	12.12
N	226	226	227	227	227
Median	170.5	971	2	21	1
Maximum	20000	555889	72	697	181
Minimum	0	0	0	0	0

Table 6.6 Descriptive statistics: Categorical variables

	Count	Percent
Processing Status		
Processed	175	77.09
Unprocessed	52	22.90
Total	227	100.00
Reuse status		
Reused	32	14.10
Not reused	195	85.90
Total	227	100.00
Multidisciplinarity		
Single discipline	177	77.97
Multidisciplinary	46	20.26
Total	223	100.00

6.5 Findings

This section summarizes the findings from the models I developed. The findings are divided into two parts: first, the models relating to whether or not data are reused, and second, the models relating to reuse impact as measured by the metrics introduced in Chapter 5.

6.5.1 What factors influence whether or not data are reused?

The first model examined reuse over the lifetime of the dataset. What factors were related to a dataset's being reused at any point during the 10 years this study's data cover? In this model I used lifetime reuse (i.e., whether the data were reused at any point) as the outcome, with the number of variables, number of cases, single- or multidisciplinarity of the data producers, processing status, the h-index of the data producers at the time the data were released, the size of the data producers' co-authorship network, and the number of primary publications before the first instance of reuse as predictors. The sample is the 227 studies authored by individuals or groups of researchers.

Individually, processing status ($X^2(1, N = 227) = 8.25$, $p = 0.004$), h-index at the time of release ($t(225) = -3.93$, $p < 0.001$), and co-authorship network size ($t(225) = -2.56$, $p = 0.01$) are significantly related to whether a dataset is reused or not. There is no significant relationship between the number of primary publications occurring before reuse and whether a dataset is reused ($t(225) = -0.06$, $p = 0.95$).

There is a significant relationship between the number of variables and whether or not a dataset is reused ($t(224) = -4.298$, $p < 0.001$) but no relationship between the number of

observations and reuse ($t(224) = 0.036$, $p = 0.97$). Finally, there was no relationship between reuse outcome and whether a dataset was multidisciplinary ($p = 0.15$, two-tailed Fisher exact test).

I fit a Firth logistic regression using the firthlogit function in Stata. The results are displayed in Table 6.7.

Table 6.7 Logistic regression of reuse outcomes for 226 datasets

Predictor	β	SE β	p	95% confidence interval		Odds ratio
(Intercept)	-4.051	0.969	0.000	-5.951	-2.151	0.017
Number of variables	0.000	0.000	0.022	0.000	0.001	1.000
Number of cases	0.000	0.000	0.695	0.000	0.000	1.000
Single vs. multi disc.	0.441	0.565	0.435	-0.666	1.548	1.554
Processing status	1.556	0.879	0.077	-0.166	3.278	4.741
h-index (release)	0.035	0.023	0.129	-0.010	0.081	1.036
Co-authorship network	0.002	0.002	0.425	-0.002	0.006	1.002
Primary pubs before reuse	-0.009	0.010	0.358	-0.028	0.010	0.991
Overall model evaluation			**X^2**	**df**	**p**	
Likelihood ratio test			18.56	7	0.010	

The model overall is significant according to the likelihood ratio test. However, the coefficients for two of the variables, number of variables and number of cases, are very small. Both of these variables have a very wide range and a large standard deviation. In order to make their coefficients more meaningful, I re-ran the model with the number of variables scaled by 100 and the number of cases scaled by 1000 (Table 6.8).

Table 6.8 Logistic regression of reuse outcomes for 226 datasets

Predictor	β	SE β	p	95% confidence interval		Odds ratio
(Intercept)	-4.051	0.969	0.000	-5.951	-2.151	0.017
Number of variables	0.049	0.021	0.022	0.007	0.091	1.050
Number of cases	0.001	0.003	0.695	-0.005	0.008	1.001
Single vs. multi disc.	0.441	0.565	0.435	-0.666	1.548	1.554
Processing status	1.556	0.879	0.077	-0.166	3.278	4.741
h-index (release)	0.035	0.023	0.129	-0.010	0.081	1.036
Co-authorship network	0.002	0.002	0.425	-0.002	0.006	1.002
Primary pubs before reuse	-0.009	0.010	0.358	-0.028	0.010	0.991
Overall model evaluation			**X^2**	**df**	**p**	
Likelihood ratio test			18.56	7	0.010	

Using the linktest function in Stata to assess the model specification shows that this is a reasonably good model for the data (Table 6.9).

Table 6.9 Linktest of 7-predictor model for reuse outcomes

	β	SE β	p	95% confidence interval	
_hat	0.758	0.339	0.026	0.927	1.423
_hatsq	-0.105	0.128	0.414	-0.355	0.147
(intercept)	-0.106	0.387	0.785	-0.863	0.652

The linktest function rebuilds the fitted model using the linear predicted value (_hat) and that value squared (_hatsq) as the predictors. In a properly specified model (i.e., all the relevant predictors are included, and all the included predictors are relevant), _hat should be significant and _hatsq should not be. In this case, _hatsq is not significant, and _hat is, which means that the model is properly specified.

However these is a potential problem with my predictors. The seven predictors used here all describe conditions at the time of data release or after. Some of the data in the sample were actually reused prior to ICPSR's release, either because they were available through

another repository, or the producers provided direct access to some researchers. To account for this, I fit another model with an additional predictor indicating whether the data had been reused prior to release, with reuse after release as the outcome variable (Table 6.10).

Table 6.10 Logistic regression of reuse outcomes for 226 datasets, including reuse before ICPSR release as a predictor

Predictor	β	SE β	p	95% confidence interval		Odds ratio
(Intercept)	-3.978	1.000	0.000	-5.938	-2.018	0.019
Number of variables*	0.047	0.000	0.039	0.002	0.910	1.048
Number of cases	0.001	0.004	0.763	-0.006	0.009	1.001
Single vs. multi disc.	0.357	0.634	0.573	-0.886	1.600	1.429
Processing status	1.145	0.897	0.202	-0.612	2.903	3.144
h-index (release)	0.037	0.024	0.131	-0.011	0.084	1.037
Co-authorship network	0.002	0.002	0.405	-0.002	0.006	1.002
Primary pubs before reuse	-0.007	0.010	0.507	-0.026	0.013	0.994
Reused prior to ICPSR release*	1.636	0.705	0.020	0.255	3.017	5.134
Overall model evaluation			X^2	**df**	**p**	
Likelihood ratio test			20.98	8	0.007	
Throughout this dissertation, *** indicates significance at the 0.001 level; ** at the 0.01 level and * at the 0.05 level.						

Again, the model is significant and the linktest also shows good results (Table 6.11).

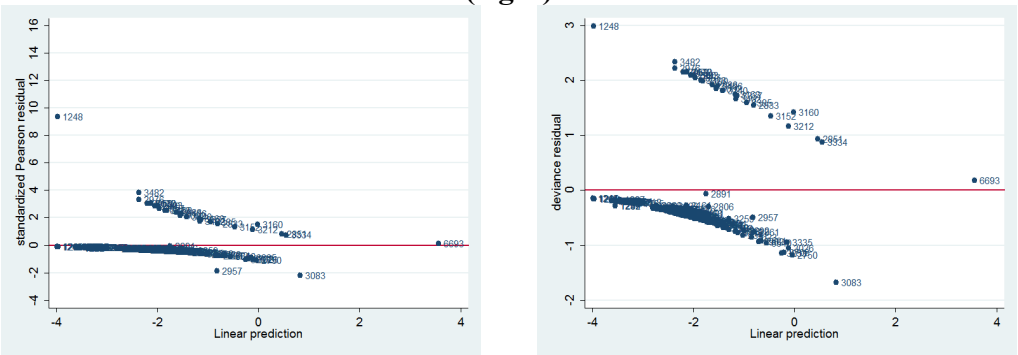
Table 6.11 Linktest on 8-predictor model for reuse outcomes

	β	SE β	p	95% confidence interval	
_hat	0.736	0.226	0.001	0.293	1.180
_hatsq	-0.107	0.085	0.210	-0.274	0.060
(intercept)	-0.119	0.402	0.767	-0.908	0.669

_hat is significant, indicating that these eight predictors are meaningful, and the non-significant p-value for _hatsq shows that there are no missing predictors.

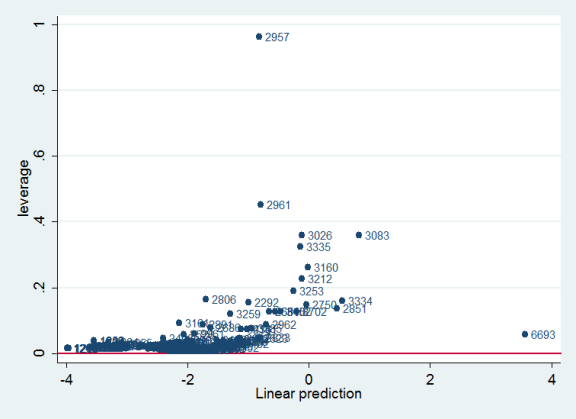
To check for influential observations, I produced plots of the standardized Pearson residuals and the deviance residuals (Figure 6.1).

Figure 6.1 Plots of standardized Pearson residuals (left) and deviance residuals (right)



Study #1248 has high residuals. However this should not have much of an effect on the model, since this study has low leverage (about 0.016). Figure 6.2 shows plots of leverage for each study in the sample.

Figure 6.2 Plot of leverage for each study



Studies #2957 and #2961 had high leverage, 0.963 and 0.453 respectively. Study #2957 was the unusual case I highlighted in Chapter 4: this study, the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment (SUPPORT), was a large study that did not produce the anticipated results. As a result, it had a very large number of primary publications (181) but no reuse. Study #2961 was also an outlier: it had an extremely large data producer co-authorship network (697) as a result of its having the largest group of data producers (7 individuals) among the datasets in my sample. These two datasets were different than the rest of the sample, and they had an outsize impact on the model. Removing these points, I re-ran the regression to arrive at the final model (Table 6.12).

Table 6.12 Logistic regression of reuse outcomes for 224 datasets

Predictor	**β**	**SE β**	**p**	**95% confidence interval**		**Odds ratio**
(Intercept)	-3.954	0.996	0.000	-5.906	-2.002	0.019
Number of variables	0.034	0.024	0.150	-0.012	0.081	1.035
Number of cases	0.002	0.004	0.688	-0.006	0.009	1.002
Single vs. multi disc.	0.202	0.643	0.754	-1.059	1.462	1.224
Processing status	1.080	0.898	0.229	-0.680	2.841	2.946
h-index (release)	0.028	0.024	0.244	-0.019	0.074	1.028
Co-authorship network	0.004	0.003	0.200	-0.002	0.009	1.004
Primary pubs before reuse	0.118	0.076	0.124	-0.032	0.267	1.125
Reuse before ICPSR release**	1.804	0.712	0.011	0.409	3.199	6.075
Overall model evaluation			**X^2**	**df**	**p**	
Likelihood ratio test			24.57	8	0.002	

The model is significant and shows good fit. The new predictor, reuse prior to release, is the only significant element in the 8-predictor model. Studies that had been reused before becoming available through ICPSR are around six times more likely to be reused after release than studies that were not. Reuse prior to ICPSR release was associated with

higher h-indices at the time of release ($t(225) = -4.09$; $p < 0.001$) but there were no significant relationships with any other variables in the model. Studies reused prior did not have different numbers of variables ($t(224) = -1.01$, $p = 0.31$); numbers of observations ($t(224) = -0.26$, $p = 0.78$); co-authorship network sizes ($t(225) = -1.44$, $p = 0.15$); or primary publications before reuse ($t(225) = 0.50$, $p = 0.61$). While they were not processed by ICPSR before they were reused, roughly the same proportion of studies reused prior to their ICPSR release were processed as those that were reused after ($X^2(N = 227) = 3.11$, $p = 0.08$). Overall, the only difference between studies reused before their ICPSR release and those first reused after is the h-index of their producers. The increase in h-index for the data producers at the time of release may be a direct result of their data's early reuse: if the reusers cited the data producers' papers (which my findings in Chapter 4 suggest is likely), those citations would have increased the data producers' h-indices.

I removed the datasets that had been reused prior to their release from the sample to test the original model on only data that had been reused after ICPSR made them available (Table 6.13).

Table 6.13 Logistic regression of reuse outcomes for datasets not reused prior to IPCSR release (*N* = 214)

Predictor	β	SE β	p	95% confidence interval		Odds ratio
(Intercept)	-3.885	0.996	0.000	-5.837	-1.932	0.021
Number of variables*	0.050	0.024	0.038	0.003	0.097	1.051
Number of cases	0.002	0.004	0.573	-0.005	0.009	1.002
Single vs. multi disc.	0.121	0.651	0.853	-1.155	1.396	1.128
Processing status	0.996	0.902	0.270	-0.772	2.763	2.706
h-index (release)	0.017	0.025	0.507	-0.032	0.065	1.017
Co-authorship network	0.004	0.003	0.193	-0.002	0.010	1.004
Primary pubs before reuse	0.142	0.086	0.100	-0.027	0.312	1.153
Overall model evaluation			**X^2**	**df**	**p**	
Likelihood ratio test			19.98	7	0.006	

The model is significant overall, shows good fit, and is well specified. For datasets that were not reused prior to being made available by ICPSR, the only significant predictor of their future reuse is how many variables the dataset contains. An increase of 100 variables results in about a 5% increase in the odds of reuse. Processing status, the number of primary publications before reuse and multidisciplinarity all have larger odds ratios that that of the number of variables, though these predictors are not significant.

Among datasets reused for the first time after their ICPSR release, there is a very strong relationship between the number of variables and the processing status of a dataset ($t(214)=-5.30$, $p < 0.001$). All datasets with more than 200 variables were processed; of the 165 processed studies in that group, 69 (42%) had fewer than 200 variables.

I recoded the number of variables as a factor with three levels comparing the top and bottom quartile to the middle 50% of the distribution. I assigned studies with fewer than

44 variables to the 'low' group, from 55-587 variables to the 'mid' group and 588 and above to the 'high' group. Table 6.14 shows the regression results.

Table 6.14 Logistic regression of reuse outcomes for datasets not reused prior to ICPSR release (*N* = 214)

Predictor	β	SE β	p	95% confidence interval		Odds ratio
(Intercept)	-4.457	1.184	0.000	-6.779	-2.136	0.012
Number of variables: Mid	-0.037	0.984	0.970	-1.967	1.892	0.964
Number of variables: High	0.590	1.005	0.557	-1.381	2.560	1.803
Number of cases	0.000	0.000	0.794	0.000	0.000	1.000
Single vs. multi disc.	0.268	0.690	0.698	-1.085	1.621	1.307
Processing status	1.489	1.285	0.247	-1.030	4.008	4.434
h-index (release)	0.008	0.030	0.783	-0.050	0.067	1.008
Co-authorship network	0.005	0.003	0.154	-0.002	0.012	1.005
Primary pubs before reuse*	0.194	0.092	0.035	0.014	0.375	1.215
Overall model evaluation			X^2	**df**	**p**	
Likelihood ratio test			21.60	8	0.006	

In this model, a previously insignificant predictor became significant: the number of primary publications before reuse. In this model, each additional primary publication raised the odds of a dataset's being reused after its ICPSR release by about 22%.

I recoded the primary publications variable into a factor to determine if there is a 'tipping point' at which reuse becomes much more likely. The table below shows the regression results.

Table 6.15 Logistic regression of reuse outcomes for datasets not reused prior to ICPSR release (*N* = 214)

Predictor	β	SE β	p	95% confidence interval		Odds ratio
(Intercept)	-4.334	1.189	0.000	-6.664	-2.003	0.013
Number of variables: Mid	-0.076	0.993	0.939	-2.022	1.871	0.927
Number of variables: High	0.626	1.010	0.535	-1.353	2.606	1.871
Number of cases	0.000	0.000	0.828	0.000	0.000	1.000
Single vs. multi disc.	0.296	0.692	0.669	-1.061	1.653	1.345
Processing status	1.442	1.293	0.265	-1.092	3.975	4.229
h-index (release)	0.015	0.028	0.594	-0.041	0.071	1.015
Co-authorship network	0.004	0.003	0.209	-0.002	0.011	1.004
3+ primary pubs before reuse*	1.149	0.577	0.046	0.018	2.280	3.156
Overall model evaluation			**X^2**	**df**	**p**	
Likelihood ratio test			19.92	8	0.010	

In this model, having three or more primary publications significantly increases the odds of a dataset's reuse by a factor of about 3. This was the most significant difference among multiple trials: there were no significant differences between having one or more publications or two or more publications compared to zero. The odds of reuse for datasets with one or two primary publications were no different than the odds of reuse for datasets with zero primary publications.

Together, these models show that the most important predictor of reuse, in general, is whether a dataset has been reused previously. Already-reused datasets are more likely to be reused again within a given time period than those that have not been reused. For datasets that have not been reused, there are two factors that are important. The number of variables a dataset contains was a significant predictor of whether or not the data would be reused; though, a small increase in the number of variables results in a small increase in the odds of reuse. For datasets within a general size category (small, medium

or large), an important predictor is how much the data producers have written about the data. In particular, datasets with three or more primary publications were more likely to be reused than those with fewer publications.

6.5.2 What factors influence reuse impact?

The above models deal with what makes a dataset likely to be reused or not, but they do not model how much reuse a dataset gets or the impact of that reuse. In this section, I draw on the impact metrics developed in Chapter 5 as outcome variables.

Prior to reusing data, potential data reusers must acquire the data; in the case of the data in my sample, they accomplish this by downloading it. To model downloaders, I used a negative binomial model. Table 6.16 lists the results.

Table 6.16 Negative binomial regression of downloaders for 227 datasets

Predictor	β	SE β	p	95% confidence interval		Incident rate ratio
(Intercept)	3.860	0.139	0.000	3.587	4.133	47.444
Number of variables*	0.015	0.007	0.041	0.001	0.029	1.015
Number of cases	0.000	0.001	0.828	-0.002	0.002	0.999
Single vs. multi disc.	0.108	0.123	0.381	-0.133	0.349	1.114
Processing status***	1.710	0.127	0.000	1.461	1.958	5.527
h-index (release)**	0.025	0.008	0.003	0.009	0.041	1.025
Coauthor network size	-0.001	0.001	0.149	-0.003	0.000	0.999
Primary publications	0.001	0.006	0.891	-0.010	0.012	1.001
Reuse prior to ICPSR release	0.497	0.258	0.054	-0.009	1.004	1.644
Overall model evaluation			X^2	**df**	**p**	
Likelihood ratio test			169.47	8	0.000	
Likelihood ratio test of alpha = 0			3.4e04	1	0.000	

The model is significant, according to a likelihood ratio test. The likelihood ratio test of alpha compares this model to a Poisson distribution model; the significance of this test shows that the negative binomial model is a better fit than a standard Poisson model.

Processing status is the most significant predictor of the number of downloaders a dataset receives, followed by the h-index of the data producer. Processed data are downloaded at a rate about 5.5 times higher than unprocessed. H-index has a marginal effect: a one-unit increase in h-index at the time of data release results in about a 2% increase in the rate at which the data are downloaded. Also significant was the number of variables, though increasing the number of variables by one has barely any effect on the rate at which data are downloaded. Prior reuse is borderline significant, but has a fairly large effect: processed studies are downloaded about 64% more than unprocessed studies.

Repeating the regression with the number of variables categorized into three groups (as before, comparing the top and bottom 25% to the middle 50% of the distribution) yields the following:

Table 6.17 Negative binomial regression of downloaders for 227 datasets

Predictor	β	SE β	p	95% confidence interval		Incident rate ratio
(Intercept)	3.795	0.138	0.000	3.524	4.067	44.497
Number of variables: Mid**	0.494	0.185	0.007	0.132	0.857	1.639
Number of variables: High***	0.680	0.206	0.001	0.275	1.084	1.973
Number of cases	0.000	0.000	0.884	0.000	0.000	1.000
Single vs. multi disc.	0.082	0.122	0.504	-0.158	0.321	1.085
Processing status***	1.349	0.189	0.000	0.978	1.720	3.853
h-index (release)***	0.026	0.008	0.001	0.010	0.042	1.027
Coauthor network size	-0.001	0.001	0.196	-0.003	0.001	0.999
Primary publications	0.001	0.006	0.926	-0.010	0.011	1.001
Reuse prior to ICPSR release	0.465	0.258	0.071	-0.040	0.970	1.591
Overall model evaluation			**X^2**	**df**	**p**	
Likelihood ratio test			175.09	9	0.000	
Likelihood ratio test of alpha = 0			3.3e04	1	0.000	

The number of variables is, overall, a significant predictor, and the Mid and High levels are significantly different than the Low level, which is the reference level. Datasets with more than 587 variables are downloaded nearly twice as frequently as those with fewer than 44 variables; datasets in the Mid level are downloaded about 64% more. In this model, the importance of processing status is reduced, with an incident rate ratio of 3.85 compared to the previous value, 5.5. H-index at the time of release remains significant but with a marginal impact.

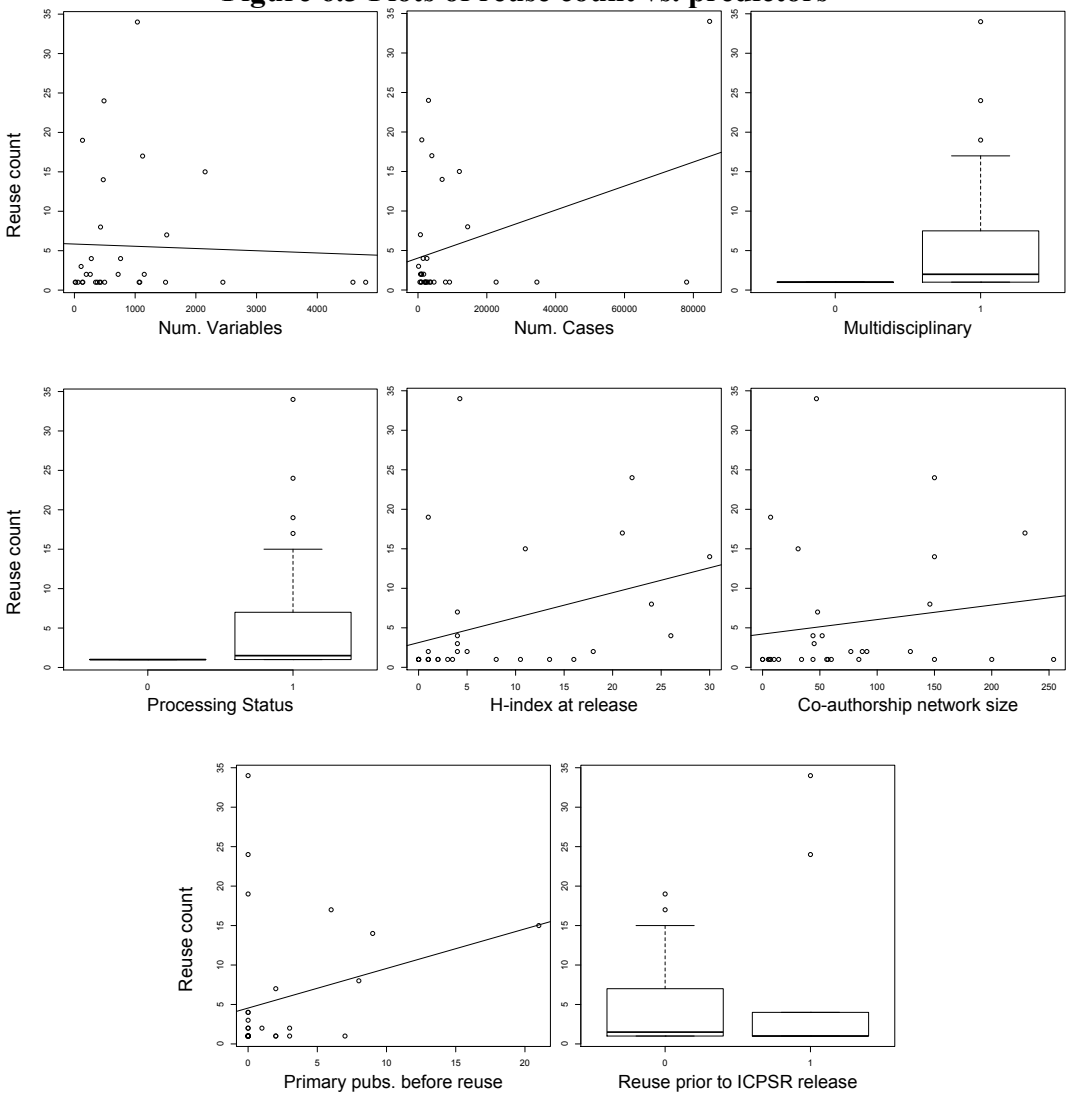
In the previous section, reuse prior to ICPSR's release was an important predictor of reuse after ICPSR release. The same does not hold true for downloads. Prior reuse approaches significance in both models ($p = 0.054$ in the model with number of variables a continuous predictor, $p = 0.072$ in the model with number of variables as a factor); in

the two models, datasets reused prior to ICPSR release are downloaded about 60-65% more than datasets without prior reuse. But in both models, that impact is secondary to the influence of processing status, and in the second model, prior reuse is also less influential than the number of variables. The downloaders model also differs from the model predicting reuse on the impact of primary publications: primary publications are not at all significant in these models.

The remaining impact measures, reuse impact, secondary impact and diversity, are only calculated for reused datasets authored by individual or group data producers ($N = 32$). This sample size is too small for regression analysis, so I explored the relationships between the predictors and the impact measures using bivariate analysis.

The first measure of impact is reuse count, how many times a dataset was used by authors other than the data producers. I eliminated one study from the analysis because it is an outlier. Study #6693, National Comorbidity Survey: Baseline (NCS-1), 1990-1992 has an exceptionally high reuse count (175) and as such is not representative of the rest of the sample. Figure 6.3 shows plots of reuse count versus the predictors individually.

Figure 6.3 Plots of reuse count vs. predictors



Visually, reuse count is positively related to all the predictors except the number of variables. However, the only significant relationship between any of the predictors and reuse count was the number of cases ($t(30) = 2.19$, $p = 0.037$), which explains about 14% of the variation in reuse count ($R^2 = 0.14$).

Table 6.18 Results of individual tests for reuse count

Predictor	*t*	*df*	*p*
Number of variables	-0.22	30	0.829
Number of cases*	2.19	30	0.037
Single vs. multi disc.	-1.20	29	0.238
Processing status	0.56	30	0.579
h-index (release)	1.97	30	0.058
Coauthor network size	0.84	30	0.405
Primary publications	1.49	30	0.148
Reuse prior to ICPSR release	-1.00	30	0.325

The previous section showed that the amount of literature, both primary and secondary, about a dataset was related to how likely the data were to be reused. The presence of secondary literature on a dataset (captured in the prior reuse variable) seems to make it more likely that someone else will reuse the data, and additionally, larger numbers of primary publications also correlate with reuse. The number of times data are reused, however, does not appear to be related to either of these factors. There is one additional aspect of literature presence not captured in these variables, though: whether or not the data producers continue using the data after others have reused them.

The data producers' publishing one or more papers after data have been reused was significantly related to reuse count ($t(30) = 4.68$, $p < 0.001$). The effect was stronger than that of the number of cases, as well, accounting for about 43% of the variation in reuse count ($R^2 = 0.43$).

Study #6693 was an outlier for secondary impact as well as for reuse count, so I eliminated it from the analysis of this metric. Secondary impact measures the impact of the papers published by data reusers; it is higher when a dataset's secondary papers are

highly cited and lower when a dataset's secondary papers are cited infrequently. In addition to the original eight predictors, I included the variable indicating whether the data producers continued to publish on the data once they had been reused.

Figure 6.4 Plot of secondary impact vs. predictors

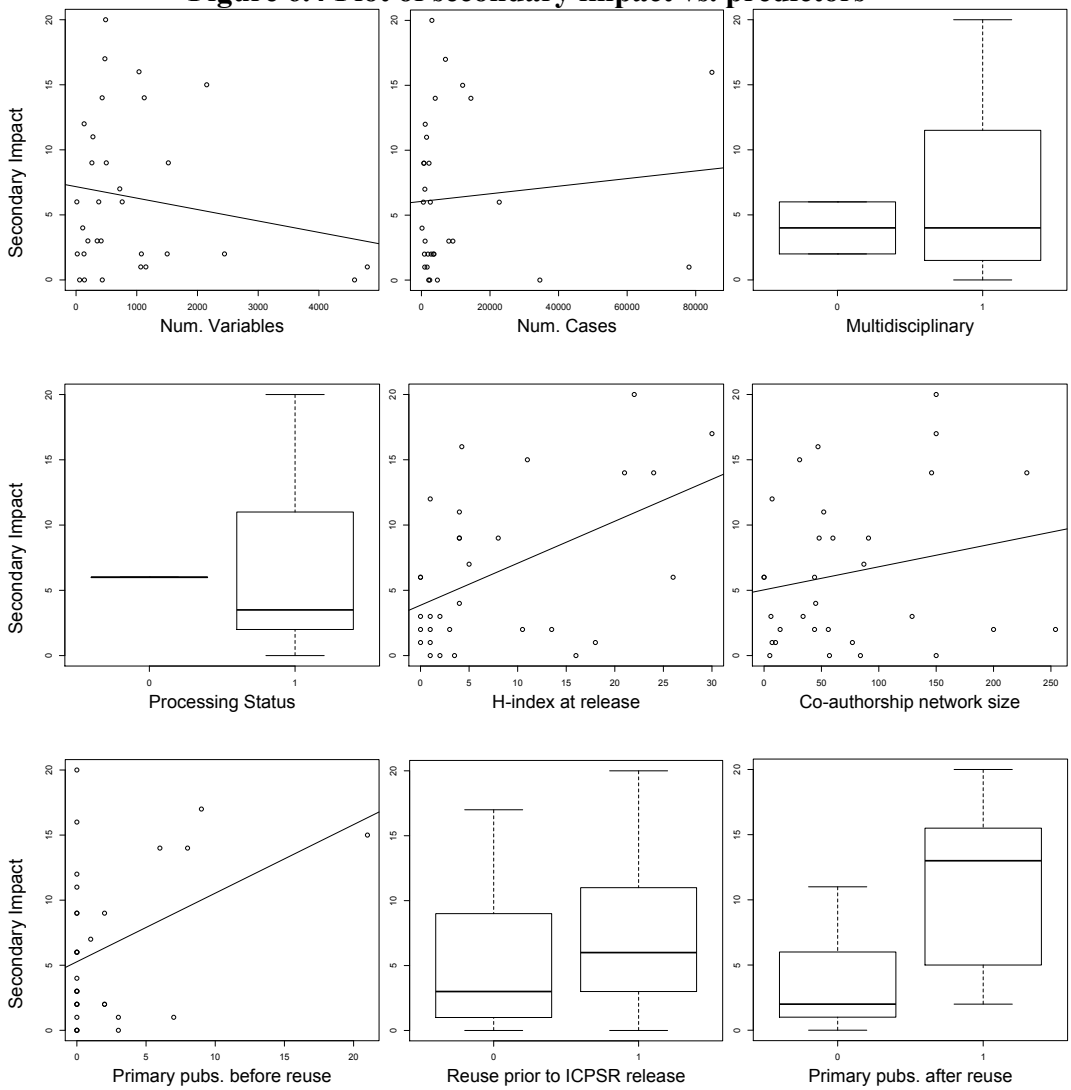


Table 6.18 shows the results of regressions of secondary impact on each of the individual predictors. Data producers' continued production of papers was the most strongly related factor to secondary impact score ($t(30) = 4.45$, $p < 0.001$), explaining about 40% of the variation in the scores ($R^2 = 0.41$). H-index at the time of release was also fairly strongly

related to secondary impact ($t(30) = 3.02$, $p = 0.005$). This factor is responsible for about a quarter of the variation in the scores on this metric ($R^2 = 0.24$). The number of primary publications before a dataset was reused was also significant ($t(30) = 2.28$, $p = 0.030$) but it had a less strong correlation ($R^2 = 0.15$).

Table 6.19 Results of individual tests for secondary impact

Predictor	*t*	*df*	*p*
Number of variables	-0.96	30	0.346
Number of cases	0.55	30	0.587
Single vs. multi disc.	0.85	29	0.401
Processing status	0.06	30	0.952
h-index (release)**	3.02	30	0.005
Coauthor network size	1.15	30	0.261
Primary publications*	2.28	30	0.030
Reuse prior to ICPSR release	1.06	30	0.296
Primary publications after reuse***	4.450	30	<0.001

Unlike secondary impact and reuse count, there were no outliers on the diversity measure, so all 32 reused studies are part of the analysis of this metric.

Figure 6.5 Plot of diversity vs. predictors

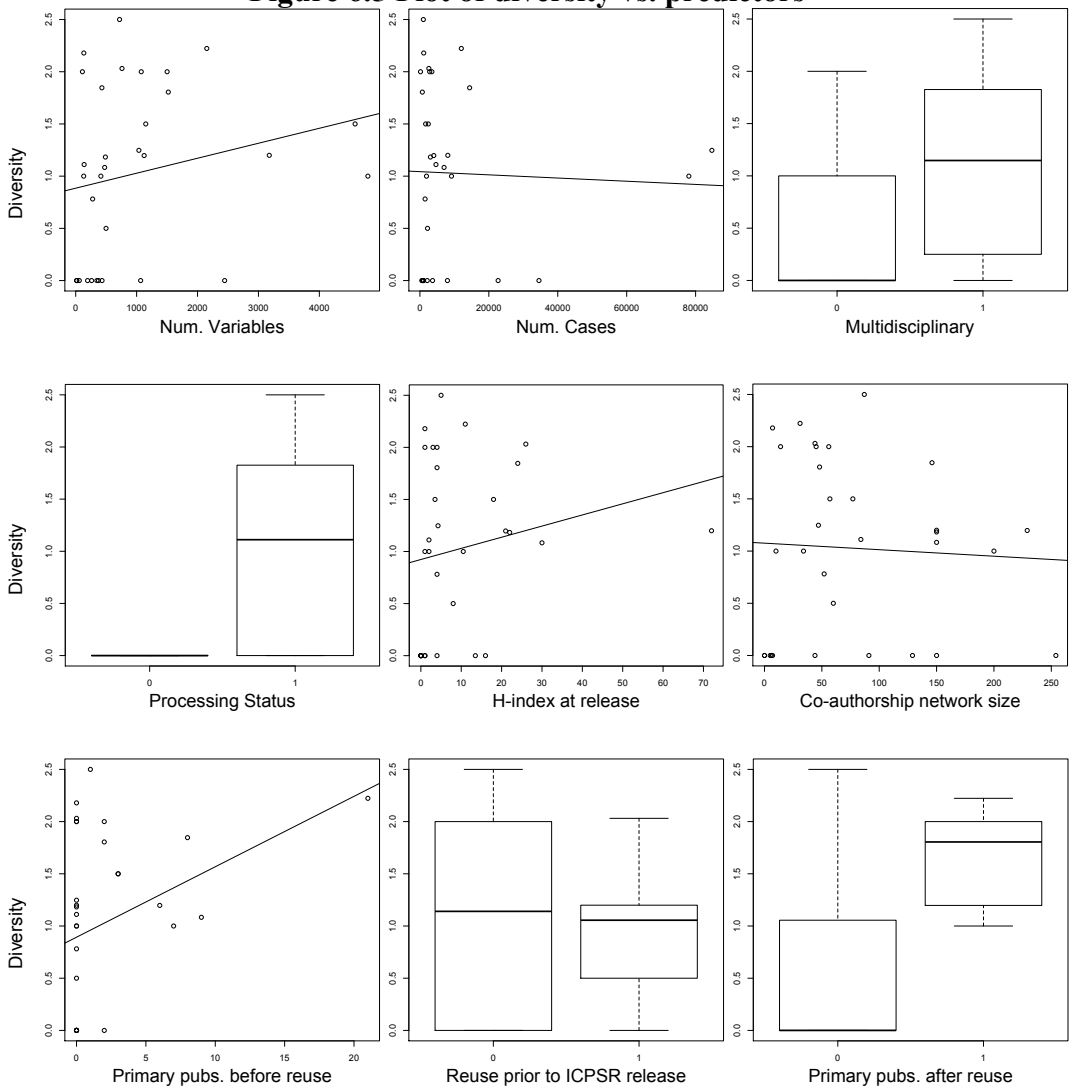


Table 6.20 Results of individual tests for diversity

Predictor	*t*	*df*	*p*
Number of variables	1.18	31	0.248
Number of cases	-0.20	31	0.841
Single vs. multi disc.	1.37	30	0.180
Processing status	1.27	31	0.216
h-index (release)	1.03	31	0.312
Coauthor network size	-0.29	31	0.777
Primary publications*	2.04	31	0.050
Reuse prior to ICPSR release	-0.55	31	0.584
Primary publications after reuse***	4.03	31	<0.001

Somewhat surprisingly, whether a dataset was single- or multidisciplinary was not related to its diversity score ($t(30) = 1.37$, $p = 0.180$). Having primary publications after reuse was strongly related to the diversity score ($t(30) = 4.03$, $p < 0.001$), with an R-squared value of 0.35. Also significant was primary publications before reuse, though that relationship was borderline ($t(31) = 2.04$, $p = 0.050$).

6.6 Discussion: Predictors of Data Reuse and Reuse Impact

In the previous chapter, I found that studies produced by government agencies or other institutions, as opposed to by individuals or groups of individuals, were much more likely to be reused. The literature on data reuse suggested a number of factors related to the decision to reuse data, but many of them do not apply to organizations, or are not measurable in a way that is comparable to metrics for individual authors. As a result, my findings in this chapter apply only to data authored by individuals or groups of individuals.

6.6.1 Research question fit (size of dataset)

The size of a dataset plays a role in whether it will be reused, how much it is downloaded and how much it is reused. Specifically, the number of variables contained in a dataset was a significant predictor in most of the models I presented above. Across the models, the effect of the number of variables was consistent. Increasing the number of variables by 100 resulted in a roughly 5% increase in the odds of reuse, and about a 2% increase in the number of downloaders. Datasets with more than 588 variables, which were in the top 25% of the distribution by variable size, were nearly twice as likely to be reused and were

downloaded twice as often as datasets in the lowest 25% of the distribution (fewer than 55 variables).

The number of cases was far less important than the number of variables. Though the number of cases was significantly related to reuse count, overall the effect of this variable was very small. Across the various models for reuse and downloaders, the odds ratio for number of cases was consistently close to one, even after I scaled the variable by 1000. Increasing the size of a dataset in terms of its number of cases appears to be less important than the number of variables.

This finding reflects the value of Big Data: larger datasets may be more flexible and better suited to a range of research questions than smaller or more narrowly focused datasets, and thus foster more scholarship. In particular, larger datasets may more readily enable research that extracts a single variable or several variables to combine with data from another source than datasets with fewer variables. These results demonstrate that broad datasets, those with many variables, are more likely to be reused than deep datasets, or those with a large number of cases.

6.6.2 Discipline of the data producer

Whether a dataset's producers represented a single or multiple disciplines was never a significant predictor of reuse or downloaders, and it was not significantly related to any of the citation-based impact metrics. In the downloaders models, the effect of discipline was small, with incident rate ratios indicating between an 8% and 11% increase in downloaders for datasets with multidisciplinary producers over those with single-

discipline producers. The effect was slightly larger in the reuse models, which showed multidisciplinarity increasing the odds of reuse by between 12% and 35%.

The relatively small effect of disciplinarity may be related to the specific repository studied here: ICPSR is first and foremost a social science data repository, and though users from all fields are welcome, ICPSR is most well known and used most heavily by social scientists. Though the different fields represented in ICPSR's collections differ to some extent in their disciplinary traditions and norms, they share an overarching social science orientation. Thus, any effect of discipline on reuse may be muted, given that the majority of users share in this orientation as well.

6.6.3 Data collection process information (processing status)

Processing status is a strong predictor of the number of downloaders, though this was not the case in any other models. However, the effect of processing status was consistently very strong. Processed studies were three to four times more likely to be reused than unprocessed studies and had more than five times as many downloaders. But whether the importance of the processing status predictor was related to documentation quality, as posited earlier in this study, is unclear. Processing status was strongly related to the number of variables in a dataset as well as to whether the data were produced by a government agency.

There is a value proposition underlying the decision to process data or not, and also, at what level the data are processed. Processed data may be more likely to be reused simply because data that is expected to be reused (especially data created by government

agencies, etc.) may be more likely to be processed. Given Niu's (2009) finding that documentation quality is not an important factor in whether data are reused or not, it is likely that processing status within this sample of datasets is more a result of an expectation of reuse rather than a determinant of reuse in and of itself. Further, processed studies are more visible within ICPSR's website. While users who know the data they want can go directly to that dataset and download it, those who are browsing rely on the tools provided by ICPSR. An important tool is variable-level searching, a search interface that allows users to find studies that contain specific questions or variables. Unprocessed studies are not indexed in a way that makes them visible in these searches, and so processing status may also have an effect on how discoverable data are.

6.6.4 Data producer reputation (h-index at time of data release)

The reputation of the data producer was not a substantial predictor of reuse. Though it was very significant in the downloaders model and significantly related to secondary impact, overall, the effect of reputation as measured by the producer's h-index was small, only around 2-3% for a one-unit increase. These results suggest that downloaders may be influenced by the reputation of the data producer if they are browsing for data without a particular target dataset in mind, but that for the most part, the reputation of the data producer is not an important part of whether or not data are reused.

6.6.5 Connection with the data producer (co-authorship network size)

Co-authorship network size was not a significant predictor of reuse or any of the reuse impact measures, nor did it ever have a meaningful effect on any of the outcome variables. The odds ratio and incident rate ratio corresponding to the co-authorship

network size predictor were consistently close to one, indicating that that predictor had very little effect on the dependent variables.

One possible explanation for this finding is that co-authorship network size was not a good measure of connection. Social networks, for example, might play a more important role than the professional networks captured through co-authorship. Another possibility is that for this sample of datasets, personal connections with the data producer are not important. This may be a difference between sharing data through repositories and sharing it informally. Personal connections with a data producer provide a way to share information about data; it may be the case that data at ICPSR are well documented enough that this type of information sharing is not necessary, or ICPSR's own support services may take the place of a connection with a data producer.

6.6.6 Prominence of data (presence in research literature)

The amount of literature about a dataset seems to be a key indicator of reuse. An increase of one primary publication prior to reuse increased the odds of reuse by between 13% and 20%, with a tipping point at three publications, at which point datasets were more than three times as likely to be reused.

In the DIPIR interviews, subjects expressed that they use existing literature to inform themselves on whether a particular dataset is any good, what can be done with it, and what has already been done with it. Though the interviewees did not strongly distinguish between secondary and primary literature, secondary literature seems to be especially important. If any secondary literature exists (i.e., the data have been reused), the odds that

a dataset will be reused again are higher. Reuse also seems to scale with the amount the data producers publish. The number of papers the data producers write does not strongly influence whether or not the data are reused, but it does influence how much they are reused. Once a dataset has been reused, the more the data producers have to say about it, the more data reusers use it as well. This is an interesting finding because it reflects the continuing value of data. If the value of these data were easily exhaustible, one would expect to see an inverse relationship between the amount the data producers publish and the amount reusers publish. Instead, within this sample, the data that producers get a lot of value from are the same data from which reusers get value.

The number of papers written by data producers is also related to the data's score on the diversity metric. The data in this sample from which data producers published multiple papers, especially when they continued to publish from their data after the data were first reused, scored higher than average on diversity. It may be the case that data that is more flexible to begin with, or is designed by the producers with many analyses in mind, generates both a large number of primary and secondary papers, and correspondingly, more diverse secondary papers. Alternatively, the more the data producers write, the more they demonstrate different approaches to take with their data, and the more information they make available to potential data reusers about the data.

The positive effect of research literature on data reuse could simply be related to publicity, in that the more often the producers put the data out into the literature, the more potential data reusers see the data and may recall it when deciding where to get new data

for a project. But the relationship between primary publications and secondary impact suggests that this is less likely. One would not expect the secondary publications to be higher-impact if their increased numbers were solely due to exposure. The importance of ongoing publishing by data producers and by reusers suggests a counterpart to Big Data: active data. Active data, regardless of size, appears to be especially valuable for supporting new, diverse and highly regarded work from existing data.

These findings provide some clarity on the relative importance of the many different factors the literature on data reuse suggest play a part in whether data are reused or not. Prior to this, numerous studies found a range of factors, from the producer reputation, to documentation quality, to a reuser's personal connection to a data producer, to contribute to data reuse. My findings show that these factors may indeed play a role – overall, the best models of data reuse outcomes included all factors – but that in the social sciences, the amount of literature, both secondary and primary, relating to a dataset is a primary predictor both of whether data will be reused and how much impact that reuse has.

CHAPTER 7

Discussion and Conclusion

Chapters 4 through 6 individually outlined findings related to my three research questions: How and why do social scientists cite data? Which datasets held by ICPSR are high-impact according to four different measures of reuse impact? What factors relate to whether data will be reused? These three questions represent distinct subsets of my guiding questions: *What is the scholarly impact of data reuse? How can stakeholders anticipate the impact the data they fund, create or curate will have?* In this chapter, I will review my findings related to the specific research questions and tie them together under the guiding question. I conclude by identifying the broader implications of this work, along with future directions.

7.1 Summary of findings

Chapter 4's analysis of data citation practices in the social sciences reviewed nearly 1,500 journal articles written by data producers and data reusers. I found that data reusers frequently do not cite the repository or source of their data, and instead more often cite the data producer's paper when describing the data they reused. However, I identified a significant increase in the number of citations to the data repository in the years 2005–2012 compared to 2000–2004. These citations typically come in conjunction with citations to the data producer's paper rather than replacing them, which suggests that reusers see different scholarly contributions in the acts of creating data (which they

acknowledge through the data producer citation) and sharing it (which they acknowledge through the repository citation). Data producers rarely indicated that their data are or would be available through a repository; only a small minority of papers included any statement of data availability, and the number did not increase over time. Noting data availability is not a priority for data producers, either because at the time they publish they do not yet know they will share the data or where they will share them, or it does not occur to them to note where the data is accessible from, or because they prefer people who want to access the data to contact them directly first.

In Chapter 5, I introduced four different measures of impact and demonstrated their use on my sample of ICPSR datasets. The four metrics address different aspects of impact. Reuse count measures the raw performance of a dataset: how many times has the dataset or some part of it been reused? Secondary impact looks at the quality of the reuse via how many times a dataset's reuse publications have themselves been cited. Diversity examines the breadth of a dataset's reuse, specifically the distribution of its reuse publication across different subject areas. Finally, the downloaders measure considers evidence of reuse separate from citations by counting the number of unique individuals who have downloaded a dataset. These metrics produced somewhat different rankings of the sample datasets, especially the diversity measure. The most reused datasets, for example, are not necessarily those that are the most broadly reused or the most downloaded. This finding shows that data can have scholarly impact in different ways and that no single indicator is the best measure of reuse impact. Instead, multiple

complementary metrics are necessary to adequately identify and understand the impact of any given dataset.

Chapter 6 turned to the question of anticipating reuse and reuse impact. I reviewed the existing literature on data reuse and analyzed a series of interviews with actual social science data reusers to compile a set of factors that might influence whether data are reused or not, and in turn, whether their reuse has high scholarly impact. The factors I identified were the reputation of the data producer, the processing status of the data, the size of the data producer's co-authorship network, the number of publications relating to the dataset, whether the data are single- or multidisciplinary, and the size of the data, in numbers of both variables and cases. In modeling the determinants of whether data were reused or not, I identified two other important factors: past success, or whether a dataset has been reused previously, and ongoing primary publication, or whether data producers continue to publish from the data after someone else has used them.

Prior reuse turned out to be the strongest predictor of whether a dataset will be reused during some later time period. When a dataset had not been reused previously, the number of primary publications was an important predictor of reuse. Datasets were far more likely to be reused after the data producers had published three or more papers from the data. Additionally, the amount of reuse is strongly related to the number of primary publications, especially primary publications that appear after a dataset is already reused once. The more the data producers continue to write, the more reuse a dataset gets, and

the more diverse that reuse is. Diversity, interestingly, was less related to whether the data were multidisciplinary than if the data producers continued to published on the data.

Downloader count, which can be related to uses other than the production of scholarly papers, was not related to the number of publications written by the data producers, but was instead associated with the data's processing status. Downloader count was also related to the size of the data, particularly the number of variables: larger datasets were downloaded more often. Larger datasets were also more likely to be reused, and to have higher reuse counts that data with smaller numbers of variables.

7.2 Implications

Though there are many stakeholders throughout the process of data production, curation, sharing, and reuse, perhaps the most important is the data producer. This study's findings show that a data producer's productivity is a key indicator of the potential reuse value of a particular dataset. Datasets that allow data producers to publish larger numbers of papers are of high value to their producers, but more than that, this study shows that they are more likely to be of value to reusers than datasets that result in fewer primary publications. Whether that connection is rooted in the possibility that data that are more suited to multiple publications from the producers are also more suited to multiple publications from others due to some kind of inherent flexibility or other characteristic, or whether primary publications somehow spur on interest from reusers is unclear, but the connection itself is important.

In some ways, this is unsurprising. Data producers are the closest to their data and thus are potentially in the best position to understand the value their data have, both for themselves and for others. When data producers feel that others will benefit from their data, they are more likely to share (Hedstrom & Niu, 2008). But how can data producers tell if their data will have that kind of value? Data producers often have a fairly narrow view of how their data might be useful. Producers' perceptions of their data's value is often closely tied to the specific research goal they have in mind, as well as where they are in the process of reaching that goal; the value of data is contingent upon local circumstance (Hilgartner & Brandt-Rauf, 1994), and so it can be difficult for scientists to image any value in their data divorced from those circumstances.

Piwowar and Chapman (2010) find journal impact factor and author experience to be positively related to data sharing. One way these factors may contribute to scientists' willingness to share is by influencing their perspective on whether their data are valuable to others. Greater experience in research might give a scientist a broader perspective on both her field and others, and as a result might expose her to more possibilities for her data. Journal impact factor can be a proxy for the overall value of the research, so scientists may feel their data that contributes to a highly placed publication is more worthwhile to share than data that do not contribute to such success. My study shows that an additional signal of value is for that scientist to look at her own productivity. If the data have been very productive and contributed to a number of primary publications, they may be likely to be so for others as well. Value to data producers is a good indicator of reuse value of all kinds.

This is a valuable insight for repositories in particular. If a dataset appears frequently in the research literature, but has not been deposited somewhere, those data may be important to recruit into a repository. Data that are published less often may be valuable to preserve for reasons other than reuse, for example because they support a seminal paper or a particularly controversial one. But given limits on space and funding for data preservation and curation that necessitate decisions about what to keep and what not to keep (Digital Preservation Coalition, 2008), the number of publications put out by the data producers could be a key selection criteria. Though basing curation decisions solely on prospective evaluations of data could result in valuable data being overlooked, this study's finding that publication activity is an indicator of reuse value provides a useful heuristic to supplement the selection processes repositories already engage in.

A major criterion for selecting data for preservation and archiving at ICPSR and at the UK Data Archive (UKDA), both major social science data repositories, is that the data "demonstrate importance to the social science community as determined by substantive value, enduring archival value, and uniqueness" (Gutmann et al., 2004, p. 213). Assessment of data producers' publication output speaks to these factors, as well as to another important consideration: the amount of metadata the data producers are able to provide when they deposit their data. For both ICPSR and the UKDA, the amount of metadata about a study already available or that can be generated before data are disseminated is key. Without appropriate metadata (the amount of which Gutman et al. characterize as "immense"), the data may be unusable. Especially important is metadata

relating to how the data were generated (i.e., the questions that were asked, the underlying research question, the research design) and processed during the course of the research and analysis (i.e., information about weighting, new variable that were generated from the data, etc.). Other studies of fields outside the social sciences have confirmed the importance of such information to data reusers (Faniel & Jacobsen, 2010; Wallis et al., 2007; Zimmerman, 2008). But in an apparently contrary finding, Niu (2009) showed that documentation quality was not especially important to social scientists' data reuse: more complete documentation was not a deciding factor in reuse.

This study showed a similar pattern. ICPSR's processed studies have more extensive metadata than those that are not processed, including more details about sampling procedures and methodology. However, whether a study was processed or not did not affect whether that study was reused or how much. Instead, primary publications had a larger influence. In prior studies, subjects expressed a need for information about how the data were produced and analyzed (Faniel & Jacobsen, 2010; Zimmerman, 2008). Though this information is present in documentation, it also exists as part of primary publications. As a result, primary publications may in part negate or surpass the influence of documentation quality. Though the information may not be aggregated or organized as well as it might be in a formal codebook, these publications may fill the need for contextual metadata, reducing the importance of the information data producers submit themselves and which the repository disseminates on their behalf. In the interviews I analyzed, several subjects noted that they would read everything they could about a study before reusing its data, so they may be using publications as their primary source of

information, relying on ICPSR's documentation only as a backup or a quick reference. Processing status is important, though, in the number of downloaders accessing a dataset. This likely reflects the fact that not all downloaders plan to use the data to produce scholarly publications, and so they may not be engaging in the same kind of background research reusers seem to do. Thus, the 'quick reference' nature of the information included with processed studies may be more important for those users.

This finding is especially interesting given the amount of work it takes for a repository to process data for reuse. For many repositories, their primary designated community is scholars, and they curate and maintain their data to be accessible and useful to researchers. But this analysis suggests that the main beneficiaries of their curation work may not be researchers, but rather students or other casual users of data. There is a high payoff for researchers from the bibliographic work ICPSR does, but they do not appear to benefit as much from ICPSR's extensive processing of data as other users might.

The connection between primary publications and reuse publications is an important communication tool and may be an effective way to encourage data producers to deposit their data. Scientists are generally more willing to share data that have already been published (Borgman et al., 2007), though some express a desire to extract as much value as possible from the data before sharing them (Tucker, 2009). But sharing data can attract more value for the producer. Scientists who have published extensively from their data are especially well positioned to reap the benefits of data sharing, since their data may be more likely to be reused often and effectively.

If increasing the number of primary publications increases the number of secondary publications as well, this could be interpreted as a reason to delay data sharing. Scientists may want to hold on to their data until they publish as many primary publications as they can, thereby maximizing the odds that their data will be reused. But in the social sciences, one study found that reused data had more total primary publications as well as secondary publications (Pienta et al., 2010); that finding aligns with this study's finding that above average reuse count is related to higher numbers of primary publications, particularly primary publications that appear after the data have been reused once.

Sharing the data does not appear to inhibit the overall number of primary publications the data producers put out, and it increases the number of secondary publications, suggesting that authors do not have anything to lose by publishing their data early. At the same time, the datasets in my sample did not show any evidence of having an 'expiration date' of any kind. New datasets were reused at a steady rate over all ten years of observation, and there were no strong indicators of when a dataset would receive its first reuse. This could indicate that small delays in data sharing are not necessarily harmful. Though there are other reasons for depositing data early on (e.g., so that the data are available when a paper is published in case someone wants to double check or replicate the work), data reuse may not be impeded much or at all if authors choose to withhold data until they successfully publish more than once.

Sharing data makes a scholarly contribution, particularly when others reuse that data to create new knowledge. But credit for making the data available is not accessible to data producers if those contributions are not visible. This study demonstrated several different metrics for making data reuse impact visible, through counting the amount of reuse, as well as its quality and breadth. Scientists are concerned with maximizing the amount of value they get from a dataset, traditionally through publishing as much as they can. But this study shows that with appropriate metrics in place, scientists can maximize the value they get from data not only by publishing on their own but by sharing.

Studies have shown mixed results on the effectiveness of direct outreach by repositories or individuals to solicit data from data producers. In the most successful study, the authors were able to obtain 59% of the data they asked for, under the guise of conducting a meta-analysis of population differential studies to assess the repeatability of result produced using a specific program (Vines et al., 2013). Other studies found less willingness among data producers to share their data. After 6 months of effort by the study team, only 26% of the authors of 141 psychology studies shared their data (Wicherts, Borsboom, Kats, & Molenaar, 2006). Even authors who published their data in an open access journal with a data sharing policy were reluctant to share in a 2009 study of papers in PLoS Medicine and PLoS Clinical Trials (Savage & Vickers, 2009). Over time, rates appear to have risen somewhat: a prior (1999) study of the same field Vines et al. examined was able to acquire data from only 1 of 30 studies (Leberg & Neigel, 1999).

These rates could potentially be increased if repositories can make a concrete, targeted value proposition to data producers. Appeals to the values of openness, sharing, and reproducibility of research seem to go only so far in incentivizing data sharing. But the argument that depositing data can result in an increase in the number of primary publications (Pienta et al., 2010), as well as attribution of credit for any secondary use, could help scientists see the tangible benefits of sharing their data. This study shows that data producers can get more value by sharing their data than withholding it.

As a practical matter, though, the ability for scientists to get credit for the contribution they make when they share their data hinges on the visibility of that reuse, particularly in terms of citations to the data. In this study, I confirm prior studies of data citation practices in the social sciences (Mooney, 2011; Sieber & Trumbo, 1995) which showed that data citation is generally lacking. That said, though formal data citations of the type recommended by ICPSR and other repositories are not as common as they ought to be, authors do generally attribute the data they reuse. But they accomplish that attribution through a citation to the data producer's paper, where the data producer lays out (typically) the rationale for the study, the methods and methodology, and other details about the data creation and analysis process.

If reusers distinguish between the contributions of the data producer and the data provider, it makes sense to encourage data reusers to include both citations to the producer's paper and to the repository from which they accessed the data. Citing the repository is crucial because it can enable automated tracking of reuse; unlike citations to

a data producer's paper, repository citations are an unambiguous indicator of reuse, since repository citations are rarely used in any other context. These citations are especially crucial for smaller datasets, which are the focus of this study.

The issue of tracking reuse touches another concern, which has not been addressed yet in this study: Collecting and managing bibliographic information is an intensive activity. ICPSR expends considerable effort in identifying publications relating to their data, but there is an important upside to that effort. Data reuse metrics are a source of value-added that repositories could offer to the scientists who deposit data with them. This study shows that this data citation information is of very high value to both producers and reusers. For reusers, bibliographic information, especially about primary publications, is an important source of contextual metadata about a study, for some even more important than the documentation that accompanies the data. For producers, bibliographic information about secondary publications is a key source of data for calculating data impact metrics.

Companies like Thompson-Reuters and Elsevier expend significant resources indexing academic literature, and part of what they gain from that expenditure is the ability to produce proprietary impact metrics. Thompson-Reuters recently debuted the Data Citation Index[14], a product that links publications to data. This database would be a natural location for implementing data impact metrics.

[14] http://wokinfo.com/products_tools/multidisciplinary/dci/

An advantage of the data metrics presented in this study, though, is that they are relatively easy to compute without extensive behind-the-scenes access to proprietary citation data. This is part of why h-index had such speedy and widespread uptake. Though different citation databases provide their own h-indices (which frequently differ from one another, due to differences in the bodies of literature they cover), h-index is calculable by anyone simply through basic access to a citation database like Web of Science (Batista, Campiteli, Kinouchi, & Martinez, 2005; Hirsch, 2005). Similarly, data metrics are primarily calculated with easily available information from Scopus. Repositories could tailor the information displayed to potential reusers based on their understanding of their designated communities.

Repositories already add value in many ways, including but not limited to providing a permanent home for data; enforcing embargoes; maintaining privacy and confidentiality of participants; and ensuring interoperability, usability and exchange of data (Albright & Lyle, 2010). Incorporating impact metrics into their suites of services could be an important additional way to add value for both data producers and reusers. Repositories could use metrics as well to gauge the strengths of their collections. How much are the data reused? Is the data in the repository used broadly or narrowly, on average? This information could be displayed as part of the study information. ICPSR already makes information about study usage available, and this study shows that prior reuse is a determinant of future reuse—knowing how, where, and by whom the data have been used previously may help potential reusers make decisions about the data they want to use.

7.3 Problems and limitations

The most important limitation of this study arises from one of the problems I addressed earlier: the lack of consistent and reliable data citation. Though ICPSR works hard to identify all literature related to their data, it is unlikely that the bibliography of data-related publications completely captures the set of publications relevant to this study. Overall, the absolute numbers of publications related to most of the datasets in my sample were low: five or fewer total publications per dataset. This means that one or two missing publications could make a big difference, particularly in how a dataset scored on the reuse impact measures. The missing publication problem is compounded when using Scopus to find additional publication information; Scopus, while it has good coverage of the social science literature, is not perfect, and it may not have provided complete numbers of citations to secondary publications, potentially skewing the secondary impact measure. With respect to the modeling work I did in Chapter 6, incomplete publication data could mean that the models I developed, while they describe the data well, do not generalize well.

Another limitation emerged from the finding that reuse was relatively rare within this sample of datasets. Just 44 out of 273 datasets in the sample had been reused, and only 31 of those were produced by individual or group authors, as opposed to governmental or institutional authors. This was problematic for the modeling section. I originally planned to explore the relationship between the predictors of reuse that I identified and datasets' scores on the impact metrics I developed, but I had too small a sample for most kinds of regression analysis. I used bivariate methods to explore relationships between the

individual factors and impact measures, but these methods cannot account for any interactions the factors may have.

On one hand, drawing my sample of datasets from studies released between 2000 and 2002 was a positive choice because it meant I could look at reuse over a long period of time with 10 years worth of citation data for each of the three yearly cohorts. At the same time, though, 10 years is a long time, and the world of data sharing and reuse may have changed significantly in that time. In 2000, direct download of ICPSR data was new. It is possible that such ready access to data may have changed how potential reusers make decisions about what data they want to reuse and how. Primary publications were important in determining whether and how much the datasets in my sample would be reused, and the interviews with current actual data reusers suggested that this is still the case. But with so much more information available just from ICPSR's website than would have been in 2000 – 2002, it is possible that the dynamics of reuse have shifted somewhat, particularly in which factors are critical to the first instance of a dataset's reuse. Perhaps quality of documentation may play a greater role, for example, when more of that documentation is prominently placed rather than when it is primarily found in the codebook.

These limitations place the study in context: it still makes significant contributions, but additional work is required to make the findings more broadly applicable.

7.4 Directions for future research

A central premise of this study is that more and better citations to data will incentivize

data producers to share their work. While this is a reasonable inference on its face, a causal relationship between citations and sharing has not been empirically tested. One possibility for future work would be to attempt to define specifically how and if citations are really an incentive for sharing data. If citations do influence sharing, data producers may choose to deposit preferentially in repositories that require or strongly recommend data citation by reusers, or those, like ICPSR, that make it easy for reusers to include citations in their papers. By comparing deposit rates at repositories that share a disciplinary focus but differ in their requirements for how reusers cite data, it may be possible to determine whether this is the case.

An obvious next step for this study would be to repeat the analysis with a larger and more recent sample of datasets. If data citations practices are getting better over time (which this study suggests is the case, at least in the social sciences), or at least related publications are easier to find, as products like Google Scholar and the Data Citation Index do a better job covering data-related literature, more recently released datasets may have a more complete corpus of publications. More complete citation information would increase the validity and reliability of this study. But more than that, repeating the study with more recent data could show whether the factors influencing reuse have changed over time.

The issue of time is an interesting one, and one that cannot be explored fully in this study. Because of sample size limitations, it is not possible to stratify the data by time period very finely. This study did not find any relevant predictors of whether a dataset was

reused soon after it was released or later on, but with a greater sample size, one could look at data reuse year by year and see if there are any relationships between the predictors I identified here and the timing of a dataset's first reuse, for example, in the first year after it is released versus the fifth year versus the tenth.

The factors I tested were developed from data reusers' own descriptions of what was important to them when considering what to reuse, and this study confirmed that the factor data reusers highlighted most—the amount of literature about a dataset—was in fact a key predictor of reuse. But there are many kinds of publications, and an interesting follow-up to this study would be a content analysis of primary publications. Having three publications seems to be an important benchmark, but is the key piece just the volume of publication, or does the type count? Primary publications can be methodology papers, presentations of preliminary results, or full reports at the end of a project. This study cannot disambiguate whether any particular type of publication has a special influence on whether the data are reused. A further possibility related to publications is the influence of their stature on reuse: are datasets related to highly-regarded publications (as measured, for example, by the impact factor of journals in which they are published) more or less likely to be reused? A deeper exploration of how data are reused and written about could yield a clearer picture of what makes data likely to be reused. The number of variables, for example, was an important predictor of reuse, but are there a few key variables or types of variables that are especially highly reused?

Another interesting route to explore would be to conduct a similar analysis in a different

field. To what extent do the patterns of citation and reuse I identified in the social sciences apply to other disciplines? The issue of literature presence did not arise in other research on data reuse; this factor came from the ICPSR interviews. Is it the case that this factor is important only in a social science context, or is it important but overlooked in the literature on data reuse in other contexts?

This study represents an early step in developing data impact metrics, and a crucial piece moving forward with that development is to identify which metrics will resonate with the communities that may eventually use them. This study proposed metrics that spanned impact of various types. What kinds of impact are especially relevant to scientists and social scientists?

7.5 Broader impacts and conclusion

Though this study focused specifically on data citation patterns, demonstrating impact metrics, and identifying factors that influence reuse, its findings speak more broadly to the issue of incentivizing data sharing and building a culture of data sharing and reuse. This study shows that data producers get more value from data by sharing them than withholding them, so long as they can get some credit for the impact reuse of their data has. Repositories can benefit by appealing to data producers' desire to maximize the value they get out of their data; by facilitating data sharing and reuse, they make it possible for others to use the data to produce new knowledge, which has a positive reflection on the data producer. This all means data citation is critical. Data citation, particularly in a format that enable automated tracking of reuse, is the key that unlocks

the power of reuse impact metrics, which in turn are an important part of fostering a culture of data sharing.

These findings are especially relevant to data that are not collected by large, federally-funded efforts, and that are not massive on their own. Much of social science research consists of survey data collected by one or a few researchers; while there are many widely known and very valuable nationally representative datasets, many more are produced on a smaller scale. These datasets may not have the name recognition of larger studies, either of the dataset itself or its producers, and are less likely to be reused overall than larger studies. But they are not without value, and within this set of small datasets, some data emerges as having higher value for reuse than others. Recognizing and rewarding sharing and reuse of this small data is important because as King (2011) notes, these data are the most at risk: they are the data that are least well documented, least frequently shared, and least reliably connected with the publications they are related to.

The findings in this study are applicable to the social sciences, and are likely to generalize to fields that work with similar data, especially data that maintains historical value over time, such as in anthropology or archaeology. By defining a connection between value for primary use and value for secondary use, this study can help data producers in such fields better assess what value their data might have to others, as well as how best to exploit the value of their data to their own ends. The findings will also help repositories recruit data and connect with data producers, in addition to shaping ongoing curation decisions.

Appendix A

List of ICPSR studies included in sample

Study Name	Study ID
Harlem Longitudinal Study of Urban Black Youth, 1968 United States	121
National Congregations Study (NCS) United States	122
What Do the Leading Indicators Lead?	1127
FOMC in 1998: Can It Get Any Better Than This?	1210
Seasonal Production Smoothing	1211
Measuring Monetary Policy Inertia in Target Fed Funds Rate Changes	1212
United States Trade Deficit and the "New Economy"	1213
Credit Unions and the Common Bond	1214
Recent Developments in the Analysis of Monetary Policy Rules	1215
State Exports and the Asian Crisis	1216
Asian Crisis and the Exposure of Large United States Firms	1217
Improving Quantitative Studies of International Conflict: A Conjecture	1218
Evolution of Monetary Policy in Transition Economies	1219
Feeding the National Accounts	1220
What Do New-Keynesian Phillips Curves Imply for Price-Level Targeting?	1221
Agricultural Outcomes and Monetary Policy Actions: Strange Bedfellows or Kissin' Cousins?	1222
Real Output in Switzerland: New Estimates for 1913-1947	1223
Bank Competition and Concentrations: The Impact of Credit Unions	1224
1980 and 1990 Industrial Structure Measures	1225
Variations in the Diffusion of State Lottery Adoptions: How Revenue Dedication Changes Morality Politics	1226
Legislative Professionalism and Incumbent Reelection: The Development of Institutional Boundaries	1227
International Bond Markets and the Intro of the Euro	1228
Do Changes in Reserves Proxy Well for Official Intervention?	1229
History of the Asymmetric Policy Directive	1230
Thresholds for Prime Rate Changes and Tests for Symmetry	1231
Information Content of Treasury Inflation-Indexed Securities	1232
Nominal Facts and the October 1979 Policy Change	1233

Library General Information Survey (LIBGIS) I [United States]: Public Library Universe, 1973	2206
Library General Information Survey (LIBGIS) III [United States]: Public Library Universe, 1977	2207
Library General Information Survey (LIBGIS) I [United States]: Public Libraries, Fiscal Year 1974	2208
Library General Information Survey (LIBGIS) III [United States]: Public Libraries, Fiscal Year 1977-1978	2209
Public Libraries Data, 1987: [United States]	2210
Public Libraries Data, 1988: [United States]	2211
Library Networks and Cooperative Library Organizations, 1977-1978: [United States]	2224
Library General Information Survey (LIBGIS) I [United States]: Public School Libraries/Media Centers, Fall 1974	2225
Survey of Special Libraries Serving State Governments, Fiscal Year 1977: [United States]	2226
Survey of Public and Private School Libraries and Media Centers, 1985: [United States]	2227
Drug Abuse Treatment Outcome Study (DATOS), 1991-1994: [United States]	2258
Library General Information Survey (LIBGIS) [United States]: Public Library Universe, 1981	2279
SHERFACS: A Cross-Paradigm, Hierarchical, and Contextually-Sensitive International Conflict Dataset, 1937-1985	2292
Justice Systems Processing of Child Abuse and Neglect Cases in a Local Jurisdiction (County) in the United States, 1993-1994	2310
Evaluation of the Midtown Community Court in New York City, 1992-1994	2311
Three-Wave Political Socialization Panel Survey of Children in the San Francisco East Bay Area, 1968-1969	2341
Psychological Classification of Adult Male Inmates in Federal Prison in Indiana, 1986-1988	2370
United States Census of Mortality: 1850, 1860, and 1870	2526
Florida's Criminal Justice Workforce Research Information System, 1985-1996	2542
Prosecution of Domestic Violence Cases in the United States, 1993-1994	2556
Benefits and Limitations of Civil Protection Orders for Victims of Domestic Violence in Wilmington, Delaware, Denver, Colorado, and the District of Columbia, 1994-1995	2557
Increasing the Efficiency of Police Departments in Allegany County, New York, 1994-1995	2558
National Survey of Weapon-Related Experiences, Behaviors, and Concerns of High School Youth in the United States, 1996	2580
Database for Forensic Anthropology in the United States, 1962-1991	2581
Evaluation of a Local Jail Training Program in Sacramento County, California, 1994-1995	2582
Evaluation of the Impact of System-Wide Drug Testing in Multnomah County, Oregon, 1991-1992	2589

Soviet Emigre Organized Crime Networks in the United States, 1992-1995	2594
Impact of Community Policing at the Street Level: An Observational Study in Richmond, Virginia, 1992	2612
Springfield [Massachusetts] Study of Populations with Disabilities, 1993-1997	2623
Criminal Behavior of Gangs in Aurora and Denver, Colorado, and Broward County, Florida: 1993-1994	2626
Reducing Disorder, Fear, and Crime in Public Housing: Evaluation of a Drug-Crime Elimination Program in Spokane, Washington, 1992-1995	2628
Developing a Juvenile Risk Assessment Instrument for Iowa State Evaluation Capacity Building, 1994-1995	2632
Evaluation of the New York City Department of Probation's Drug Treatment Initiative, 1991-1994	2652
Bethlehem [Pennsylvania] Police Family Group Conferencing Project, 1993-1997	2679
Evaluation of the Children at Risk Program in Austin, Texas, Bridgeport, Connecticut, Memphis, Tennessee, Savannah, Georgia, and Seattle, Washington, 1993-1997	2686
Services Research Outcomes Study, 1995-1996: [United States]	2691
Work and Family Services for Law Enforcement Personnel in the United States, 1995	2696
Convenience Store Crime in Georgia, Massachusetts, Maryland, Michigan, and South Carolina, 1991-1995	2699
Assessment of a Single-Purpose Substance Abuse Facility for Committed Juvenile Offenders in Virginia, 1995-1997	2730
Survey of Disabled and Nondisabled Adults, 1972: [United States]	2731
Evaluation of Victim Services Programs Funded by “Stop Violence Against Women” Grants in the United States, 1998-1999	2735
Incapacitation Effects of Incarcerating Drug Offenders: Longitudinal Arrest Histories of Adults Arrested in Washington, DC, 1985-1986	2741
Strategies for Retaining Offenders in Mandatory Drug Treatment Programs in Kings County, New York, 1994-1995	2749
Survey of Gun Owners in the United States, 1996	2750
Evaluating Anti-Gang Legislation and Gang Prosecution Units in Clark and Washoe Counties, Nevada, 1989-1995	2753
Gambling Impact and Behavior Study, 1997-1999: [United States]	2778
National Study of Innovative and Promising Programs for Women Offenders, 1994-1995	2788
Effectiveness of Alternative Victim Assistance Service Delivery Models in the San Diego Region, 1993-1994	2789
World Values Surveys and European Values Surveys, 1981-1984, 1990-1993, and 1995-1997	2790
Children's Out-of-Court Statements: Effects of Hearsay on Jurors' Decisions in Sacramento County, California, and Maricopa County, Arizona, 1994-1997	2791
Prevalence of Five Gang Structures in 201 Cities in the United States, 1992 and 1995	2792

Manhattan [New York City], New York, 1996-1997	
Continuity and Change in Criminal Offending by California Youth Authority Parolees Released 1965-1984	3136
French National Election Study, 1997	3138
United States Presidential Election Campaign Media Analysis, 1996	3139
Role of Stalking in Domestic Violence Crime Reports Generated by the Colorado Springs Police Department, 1998	3142
CrimeMapTutorial Workbooks and Sample Data for ArcView and MapInfo, 2000	3143
Gun Density, Gun Type, and the Dallas Homicide Rate, 1980-1992	3145
Police Use of Force in Metro-Dade, Florida, and Eugene and Springfield, Oregon, 1993-1995	3152
Project on Policing Neighborhoods in Indianapolis, Indiana, and St. Petersburg, Florida, 1996-1997	3160
Valuation of Specific Crime Rates in the United States, 1980 and 1990	3161
Harvard School of Public Health College Alcohol Study, 1997	3163
Evaluation of Grants to Encourage Arrest Policies for Domestic Violence Cases in the State College, Pennsylvania, Police Department, 1999-2000	3166
Understanding the Use of Force By and Against the Police in Six Jurisdictions in the United States, 1996-1997	3172
Vital Statistics: Divorce Detail [United States], 1986	3176
Vital Statistics: Divorce Detail [United States], 1987	3178
Local Health Department Infrastructure Study, 1999-2000: [United States]	3185
Evaluating a Driving While Intoxicated (DWI) Night Drug Court in Las Cruces, New Mexico, 1997-1998	3186
Felonious Homicides of American Police Officers, 1977-1992	3187
National Organizations Survey (NOS), 1996-1997	3190
Longitudinal Study of Violence Against Women: Victimization and Perpetration Among College Students in a State-Supported University in the United States, 1990-1995	3212
Integrating the Ion Mobility Spectrometer Into Drug Monitoring at the New Orleans Pretrial Diversion Program, 1996	3213
Homicides in New York City, 1797-1999 [And Various Historical Comparison Sites]	3226
Six-State Survey of Elderly Dual Enrollees in Medicare and Medicaid, 1999	3240
Impact Evaluation of Stop Violence Against Women Grants in Dane County, Wisconsin, Hillsborough County, New Hampshire, Jackson County, Missouri, and Stark County, Ohio, 1996-2000	3252
Resources for Enhancing Alzheimer's Caregiver Health, 1996-2001	3253
Validation of a Risk Assessment Instrument for Juvenile Probationers in Alameda County, California, 1996	3254
National Evaluation of Operation Drug Test, 1998-1999	3259
Spatial Analysis of Crime in Appalachia [United States], 1977-1996	3260
Neighborhood Revitalization and Disorder in Salt Lake City, Utah, 1993-2000	3261

States, 1998-2001	
Evaluation of Violence Against Women With Physical Disabilities in Michigan, 2000-2001	3414
Treatment of Incarcerated Women with Substance Use Disorder and Post-traumatic Stress Disorder in Providence, Rhode Island, 1999-2001	3416
Height of Free African Americans in Maryland, 1800-1864	3422
Evaluation of a Truancy Reduction Program in Nashville, Tennessee, 1998-2000	3424
Pennsylvania Sentencing Data, 1998	3450
Agricultural and Demographic Records for Rural Households in the North, 1860: [Instructional Materials]	3463
New Orleans Slave Sample, 1804-1862 [Instructional Materials]	3464
Processing and Outcome of Death Penalty Appeals After Furman v. Georgia, 1973-1995: [United States]	3468
National Congregations Study, 1998 and 2006 [Cumulative File]	3471
California Vital Statistics and Homicide Data, 1990-1999	3482
Evaluation of the Regional Auto Theft Task (RATT) Force in San Diego County, 1993-1996	3483
Evaluation of a Centralized Response to Domestic Violence by the San Diego County Sheriff's Department Domestic Violence Unit, 1998-1999	3488
Survey of Citizens' Attitudes Toward Community-Oriented Law Enforcement in Alachua County, Florida, 1996	3491
The State and Life Chances in Urban China, 1949-1994	3552
White-Collar Criminal Careers, 1976-1978: Federal Judicial Districts	6540
Effects of the Baltimore County, Maryland, Police Department's Community-Oriented Drug Enforcement (Code) Program, 1990-1992	6644
National Comorbidity Survey: Baseline (NCS-1), 1990-1992	6693
Intensive Community Supervision in Minnesota, 1990-1992: A Dual Experiment in Prison Diversion and Enhanced Supervised Release	6849
Electoral and Demographic Data for New York, 1830-1875	6926
Beliefs About Social Stratification, 1980: [United States]	8702
European Survey About Teachers and Cancer Prevention, 1989	9407
Study of Race, Crime, and Social Policy in Oakland, California, 1976-1982	9961

Appendix B

Citation Database Comparison

Below, I present a review of literature comparing commonly used citation databases.

B.1 Citation Databases

B.1.1 Web of Science (ISI)

Web of Science was first introduced in 1963 as an analog resource for citations in the sciences (Garfield, 1963). The database is comprised of several different indexes, the two most prominent of which are the Science Citation Index, which covers 8,500 journals and the Social Science Citation Index, which indexes over 3,000 social science journals. Both of these indexes include publications as far back as 1900. Web of Science also includes tools for citation analysis, including citation mapping and computation of basic metrics like h-index, times cited, average citations per item and year, etc. (Thomson Reuters, n.d.).

B.1.2 Scopus (Elsevier)

Introduced in 2004 (Reed Elsevier, 2004), as of November 2012, Scopus indexes 19,500 journals in the sciences, social sciences, and humanities, the bulk of which are from 1996 and after (Elsevier B. V., 2012). In addition to interface displays of bibliometric information like h-index, journal impact, etc., Scopus also provides API access for

automated retrieval of citation information for analysis, including author and affiliation information.

B.1.3 Google Scholar

Google Scholar was released around the same time Scopus appeared in 2004 (Payne, 2004). Google Scholar is a free search service created and maintained by Google. It indexes "most major academic publishers and repositories worldwide," though Google does not disclose exactly which publishers and repositories they index (Google, n.d.). Google Scholar's citation information can be unreliable due to duplication or errors (Falagas, Pitsouni, Malietzis, & Pappas, 2008) or even through its vulnerability to spam (Beel & Gipp, 2010). As Meho & Yang (2007) point out, Google is not transparent about how Google Scholar indexes and how often the index is updated. They suggest that until Google is more forthcoming about the processes underlying Google Scholar, that service should be consulted only as a supplement, not as a primary research tool.

Google Scholar calculates h-index for journals as well as identifying how often and by whom a publication is cited. It does not allow bulk export or programmatic access to citation information, however. Furthermore, results sets are limited to 1,000 items.

B.1.4 Microsoft Academic Search

Microsoft Academic Search is a newer, free service. It was launched in 2009 but only began indexing disciplines other than computer science in 2011. Coverage of social science journals began in September 2011. Currently, 87 content providers participate in

Academic Search, including the American Psychological Association, JSTOR, the Social Science Research Network (SSRN), among others (Microsoft, n.d.).

Academic Search provides extensive tools for visualizing citation and author networks, along with API access for users to build their own apps and tools.

B.2 Comparing the Databases

Until 2004, the only source for citation analysis was Web of Science. The introduction of new citation databases raised questions as to which resource was best suited for my bibliometric analysis. Which database has the most complete coverage, both over time and across disciplines? Which has the most accurate coverage, i.e., lacking duplicates or incorrect citations? Which has the best tools to support researchers in identifying and compiling citations?

The choice of citation database can have substantive effects on metrics calculated from them (Bar-Ilan, 2007). Those effects differ both in type and magnitude across disciplines since citation coverage varies not only between databases but between fields as well ("ISI Coverage by Discipline," 2005). Numerous researchers have addressed this issue in a wide range of fields: medicine (Falagas et al., 2008; Strotmann & Zhao, 2010), human-computer interaction (Meho & Rogers, 2008), library and information sciences (Meho & Yang, 2007), social sciences (Norris & Oppenheim, 2007), social work (Lasda Bergman, 2012), oncology, (López-Illescas, de Moya-Anegón, & Moed, 2008), and earth sciences (Mikki, 2009).

Their findings show that Scopus has more extensive coverage than Web of Science both in terms of number of journals indexed as well as at the article level; this is true for the social sciences as well as other disciplines (Norris & Oppenheim, 2007, p. 168). Google Scholar consistently provides a greater number of citations than any other database (Lasda Bergman, 2012; Meho & Yang, 2007; Mikki, 2009) and is especially valuable for identifying hard-to-find items not indexed by any other source (Falagas et al., 2008).

Concerns about coverage are tightly linked to the issue of ensuring the quality of impact metrics computed from bibliometric analyses. Studies focusing on the effect of using different citation databases as the source for impact metrics find, interestingly, that while the raw number of citations varies from database to database, journals and researchers are ranked similarly across databases (Lasda Bergman, 2012; Meho & Rogers, 2008). While Bar-Ilan's 2007 finding that the absolute value of the h-index for a researcher is different when computed based on Web of Science than Scopus data is confirmed by other research, the relative ranking of individuals does not change, irrespective of the source data used (Meho & Rogers, 2008). The same holds true for journal impact factors (Lasda Bergman, 2012). Scopus or Google Scholar may have a larger pool of items on which to calculate metrics, but the end result is largely the same as when using Web of Science to generate rankings.

This may be in part of the specifics of the coverage range of each database. López-Illescas et al. (2008) examined coverage of oncology journals within Scopus and Web of Science and found that the oncology journals in Scopus that are not covered by Web of

Science are for the most part relatively low-impact (2008, p. 311). Thus, the inclusion of citations to and from these journals has only a marginal effect on impact metrics computed from these sources. Similarly, although Google Scholar consistently produces a higher number of citations, the influence on impact metrics is small. Lasda Bergman found that when only journal articles are counted, Google Scholar's citation advantage is much diminished (2012, p. 373), and as a result, the overall impact ranking of journals in her sample was the same across citation databases (2012, p. 377).

Norris and Oppenheim conclude that Scopus is the best resource for social science citation analysis (2007). Strotmann and Zhao (2010) recommend that when dealing with interdisciplinary or multidisciplinary fields, researchers combine a commercial index like Scopus or Web of Science with an open bibliographic resource, like PubMed or Google Scholar. Based on her study of social work research, Lasda Bergman suggests that it would be best to use Web of Science, Scopus, and Google Scholar together (2012). In the end, none of the citation databases available are definitively better than the others: each has tradeoffs in terms of coverage, tools available, etc. Which is the ‘best’ citation resource depends on the discipline or field under study as well as the research question to be answered.

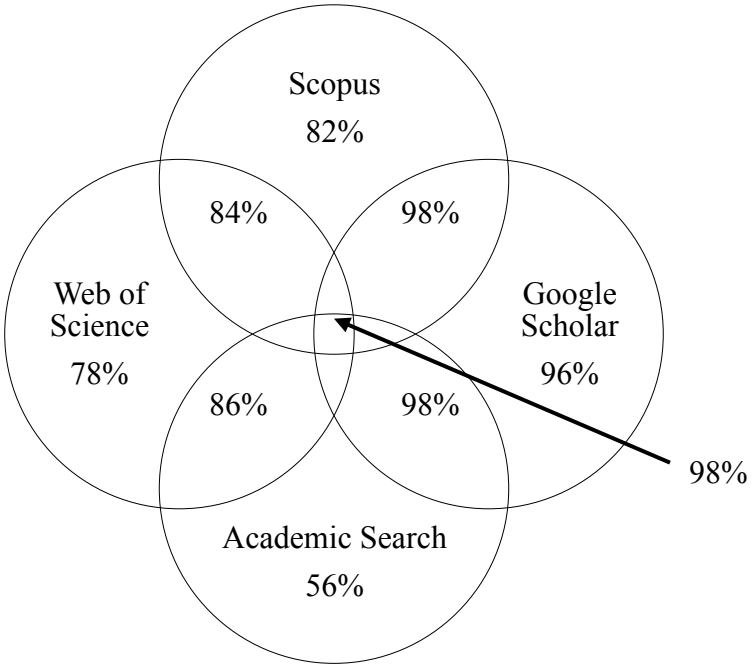
For this dissertation, the key criterion for the citation databases I used was coverage of the journals represented in my sample of reuse publications. In order to determine which citation database had the best coverage for my purposes, I took a random sample of 50 articles from the full sample of 449 reuse publications and compared the citation counts

for these articles from Scopus, Web of Science, Google Scholar, and Microsoft Academic Search.

Table B.1 Success rate comparison for finding 50 journal articles in 4 citation databases

Database	Percent of Documents Found
Google Scholar	96.00%
Scopus	82.00%
ISI Web of Science	78.00%
Microsoft Academic Search	56.00%

Figure B.1 Percentage of articles (*N* = 50) indexed by each database



Individually, Google Scholar has the greatest coverage (48 articles, 96%) and Academic Search the smallest (28 articles, 56%). Using a Cochran's Q test, I found that there is a significant difference in coverage among the four citation databases ($Q = 27.75$; $p < 0.001$). Pairwise comparisons (using a Bonferroni-corrected threshold of $p < 0.008$)

showed that Microsoft Academic Search has significantly smaller coverage than any of the other three citation databases. The most effective combination of databases is Google Scholar plus any of the other individual databases; in all cases, combining Google Scholar with another database covers 98% of the articles in the sample.

A Friedman test showed a statistically significant difference in the number of citations reported by each database ($X^2(2, N = 50) = 66.14$; $p < 0.001$). Post-hoc analysis with Wilcoxon signed-rank tests, again using a Bonferroni-corrected threshold of $p < 0.008$, further showed significant differences in the numbers of citations reported by each pair of databases. The median number of citations was 19 for Scopus, 17.5 for Web of Science, 9 for Microsoft Academic Search, and 22 for Google Scholar.

Despite differences in the numbers of citations reported by each citation database, the relative rankings of the articles were similar across them. Pair-wise comparisons of the databases using Spearman's rank order correlation showed statistically significant relationships between the rankings produced by each citation database ($p < 0.001$ in all cases).

This analysis shows that for my sample of articles, Google Scholar and Scopus have the best coverage—especially when combined—and these two databases provide the largest number of citations per article compared to Microsoft Academic Search and Web of Science. Google Scholar may have somewhat better coverage, especially for foreign language texts, but Scopus has sufficient coverage with the added advantage that it is a

better research tool. Combined, these two citation databases have close to complete coverage of my set of reuse publications. In this dissertation, I rely primarily on Scopus and the citation information it reports, supplemented by information from Google Scholar when an article is not in Scopus.

Appendix C

Sample g-index calculation for hypothetical datasets

Dataset A: 5 publications, each with 5 citations
Dataset B: 5 publications, each with 25 citations

Process (Egghe, 2006): List publications in rank order based on number of citations. A publication's g-index is the highest rank at which the cumulative citations are equal or less than the rank squared. If this condition cannot be met with the number of extant publications, publications with 0 citations are added until the g-index can be determined.

Dataset A			
Publication Rank	Number of citations	Cumulative citations	$Rank^2$
1	5	5	1
2	5	10	4
3	5	15	9
4	5	20	16
5	5	25	25
6	0	25	36

Dataset B			
Publication Rank	Number of citations	Cumulative citations	$Rank^2$
1	25	25	1
2	25	50	4
3	25	75	9
4	25	100	16
5	25	125	25
6	0	125	36
7	0	125	49
8	0	125	64
9	0	125	81
10	0	125	100
11	0	125	121
12	0	125	144

Appendix D

Impact metric scores and ranking for 44 datasets

Study ID	Reuse Count	Reuse Count Rank	Sec. Impact	Sec. Impact Rank	Down-loaders	Down-loaders Rank	Diversity	Diversity Rank
1248	1	26	6	20	268	93	0.000	41
2258	19	5	21	2	1244	12	1.453	21
2292	1	26	6	20	220	117	0.000	41
2580	2	20	9	15	569	41	1.111	28
2581	1	26	3	28	677	34	1.556	17
2686	1	26	2	32	773	26	0.500	36
2691	1	26	5	23	454	58	0.500	36
2731	1	26	3	28	118	174	0.000	41
2750	1	26	2	32	711	31	1.500	18
2778	15	8	18	5	2637	3	1.351	24
2833	17	7	14	10	1984	6	1.355	23
2835	2	20	5	23	1252	10	0.563	35
2851	24	4	20	3	818	22	1.918	5
2854	1	26	4	26	119	173	1.667	10
2857	1	26	1	38	297	83	1.111	28
2884	32	3	20	3	748	27	2.139	3
2976	19	5	12	13	1233	13	1.392	22
2993	1	26	2	32	397	62	0.500	36
3002	7	14	9	15	1079	16	1.601	13
3023	9	11	14	10	791	25	1.950	4
3030	2	20	3	28	345	71	1.333	25
3052	3	18	4	26	478	53	1.049	32
3062	2	20	2	32	267	94	1.667	10
3088	8	12	5	23	2478	4	1.589	15
3136	1	26	2	32	318	76	1.500	18
3152	4	16	11	14	215	120	1.319	26
3160	34	2	16	8	819	21	1.587	16
3163	8	12	14	10	720	30	1.706	9
3172	1	26	3	28	484	50	1.000	33
3190	2	20	7	18	882	20	2.847	1
3212	1	26	0	41	1103	15	0.000	41
3240	1	26	0	41	145	158	1.111	28
3323	1	26	2	32	242	103	0.500	36
3334	4	16	6	20	1658	7	2.587	2

3335	1	26	1	38	360	67	1.000	33
3337	15	8	15	9	1250	11	1.798	6
3355	3	18	7	18	2209	5	1.741	8
3385	14	10	17	6	600	40	1.793	7
3402	2	20	1	38	806	23	1.319	26
3450	7	14	17	6	442	59	1.667	10
3482	1	26	0	41	916	17	1.500	18
6540	1	26	—	—	231	110	1.111	28
6693	175	1	83	1	3787	1	1.593	14
8702	1	26	9	15	388	64	0.500	36

BIBLIOGRAPHY

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010). Metrics: Do metrics matter? *Nature News*, *465*(7300), 860–862. doi:10.1038/465860a

Abramo, G., D'Angelo, C. A., & Viel, F. (2013). Assessing the accuracy of the h- and g-indexes for measuring researchers' productivity. *Journal of the American Society for Information Science and Technology*, *64*(6), 1224–1234. doi:10.1002/asi.22828

Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Future impact: Predicting scientific success. *Nature*, *489*(7415), 201–202. doi:10.1038/489201a

Agosti, V., Nunes, E., & Levin, F. (2002). Rates of psychiatric comorbidity among U.S. residents with lifetime cannabis dependence. *The American journal of drug and alcohol abuse*, *28*(4), 643–652.

Akmon, D. (2011). Moving beyond sharing vs. withholding to understand how scientists share data through large-scale, open access databases. In *Proceedings of the 2011 iConference* (pp. 634–635). New York, NY, USA: ACM. doi:10.1145/1940761.1940848

Akmon, D., Zimmerman, A., Daniels, M., & Hedstrom, M. (2011). The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs. *Archival Science*, *11*(3-4), 329–348. doi:10.1007/s10502-011-9151-4

Albright, J. J., & Lyle, J. A. (2010). Data Preservation through Data Archives. *PS: Political Science & Politics*, *43*(01), 17–21. doi:10.1017/S1049096510990768

Altman, M., & King, G. (2007). A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine*, *13*(3/4). doi:10.1045/march2007-altman

Antelman, K. (2004). Do Open-Access Articles Have a Greater Research Impact? *College & Research Libraries*, *65*(5), 372–382.

Bar-Ilan, J. (2007). Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, *74*(2), 257–271. doi:10.1007/s11192-008-0216-y

Batista, P. D., Campiteli, M. G., Kinouchi, O., & Martinez, A. S. (2005). Universal Behavior of a Research Productivity Index. *arXiv:physics/0510142*. Retrieved from http://arxiv.org/abs/physics/0510142

Baudoin, P. (2008). On Palimpsest and Data. *The Serials Librarian*, *55*, 376–378.

Beagrie, N., Chruszcz, J., & Lavoie, B. (2008). *Keeping Research Data Safe*. JISC. Retrieved from http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx

Beel, J., & Gipp, B. (2010). Academic Search Engine Spam and Google Scholar's Resilience Against it. *The Journal of Electronic Publishing*, *13*(3). doi:10.3998/3336451.0013.305

Berwick, D. M. (1995). The SUPPORT Project: Lessons for Action. *Hastings Center Report*, *25*(6), S21–S22. doi:10.2307/3527853

Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: supporting sharing in science and engineering. In *Proceedings of the 2003 international ACM SIGGROUP*

conference on supporting group work (pp. 339–348). New York, NY, USA: ACM. doi:10.1145/958160.958215

Bollen, J., Van de Sompel, H., & Rodriguez, M. A. (2008). Towards usage-based impact metrics: first results from the mesur project. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries* (pp. 231–240). New York, NY, USA: ACM. doi:10.1145/1378889.1378928

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *Int. J. Digit. Libr.*, *7*(1), 17–30. doi:10.1007/s00799-007-0022-9

Bourke, P., Butler, L., Biglia, B., & Australian National University. Research Evaluation and Policy Project. (1996). *Monitoring research in the periphery: Australia and the ISI indices*. Canberra: The Project. Retrieved from http://trove.nla.gov.au/version/46724864

Bourne, P. E. (2005). Will a Biological Database be Different from a Biological Journal. *PLoS Computational Biology*, *1*, 179–181. doi:10.1371/journal.pcbi.0010034

Bowker, G. C. (2000). Biodiversity Datadiversity. *Social Studies of Science*, *30*(5), 643–683. doi:10.2307/285760

Brembs, B., & Munafò, M. (2013). Deep Impact: Unintended consequences of journal rank. *arXiv:1301.3748*. Retrieved from http://arxiv.org/abs/1301.3748

Broadus, R. N. (1971). The Literature Of The Social Sciences: A Survey Of Citation Studies. *International social science journal*, *23*(2), 236–243.

Brown, C. (2003). The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance. *Journal of the American Society for Information Science and Technology*, *54*(10), 926–938. doi:10.1002/asi.10289

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. doi:10.2218/ijdc.v7i1.218

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., & Blumenthal, D. (2002). Data withholding in academic genetics: evidence from a national survey. *JAMA: the journal of the American Medical Association*, *287*(4), 473–480. doi:10.1001/jama.287.4.473

Carlson, S., & Anderson, B. (2007). What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. *Journal of Computer-Mediated Communication*, *12*(2), 635–651. doi:10.1111/j.1083-6101.2007.00342.x

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(Suppl 15), S2. doi:10.1186/1471-2105-12-S15-S2

Committee on Responsibilities of Authorship in the Biological Sciences. (2003). *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Washington, D.C.: National Research Council. Retrieved from http://www.nap.edu/catalog.php?record_id=10613

Costas, R., & Bordons, M. (2008). Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, *77*(2), 267–288. doi:10.1007/s11192-007-1997-0

Costas, R., Meijer, I., Zahedi, Z., & Wouters, P. (2013). *The Value of Research Data: Metrics for datasets from a cultural and technical point of view*. Copenhagen, Denmark: Knowledge Exchange. Retrieved from http://www.knowledge-exchange.info/datametrics

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023 – 4038. doi:10.1098/rsta.2010.0165

Daniels, M., Faniel, I., Fear, K., & Yakel, E. (2012). Managing fixity and fluidity in data repositories. In *Proceedings of the 2012 iConference* (pp. 279–286). New York, NY, USA: ACM. doi:10.1145/2132176.2132212

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., & Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: randomised controlled trial. *BMJ*, *337*(jul31 1), a568–a568. doi:10.1136/bmj.a568

Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, *64*(5), 1055–1064. doi:10.1002/asi.22806

Digital Preservation Coalition. (2008). *Preservation Management of Digital Materials: A Handbook*. Retrieved from http://www.dpconline.org/advice/preservationhandbook

Donnelly, W. J. (1996). The SUPPORT project and improving care for seriously ill patients. *JAMA: the journal of the American Medical Association*, *275*(16), 1228; author reply 1230–1231.

Donovan, C. (2008). The Australian Research Quality Framework: A live experiment in capturing the social, economic, environmental, and cultural returns of publicly funded research. *New Directions for Evaluation*, *2008*(118), 47–60. doi:10.1002/ev.260

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690. doi:10.1177/0306312711413314

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152. doi:10.1007/s11192-006-0144-7

Ellison, G. (2010). *How Does the Market Use Citation Data? The Hirsch Index in Economics* (Working Paper No. 16419). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w16419

Elsevier B. V. (2012). Scopus In Detail. Retrieved March 8, 2013, from http://www.info.sciverse.com/scopus/scopus-in-detail/facts/

Engels, T. C. E., Ossenblok, T. L. B., & Spruyt, E. H. J. (2012). Changing publication patterns in the Social Sciences and Humanities, 2000–2009. *Scientometrics*, *93*(2), 373–390. doi:10.1007/s11192-012-0680-2

Eysenbach, G. (2011). Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. *Journal of Medical Internet Research*, *13*(4), e123. doi:10.2196/jmir.2012

Falagas, M. E., & Alexiou, V. G. (2008). The top-ten in journal impact factor manipulation. *Archivum Immunologiae et Therapiae Experimentalis*, *56*(4), 223–226. doi:10.1007/s00005-008-0024-5

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, *22*(2), 338–342. doi:10.1096/fj.07-9492LSF

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Computer Supported Cooperative Work (CSCW)*, *19*(3-4), 355–375. doi: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 American Society for Information Science and Technology*, *49*(1), 1–10. doi:10.1002/meet.14504901068

Fear, K. (2011). "You made it, you take care of it": Data Management as Personal Information Management. *International Journal of Digital Curation*, *6*(2), 53–77. doi:10.2218/ijdc.v6i2.190

Frank, R. R. (1996). The SUPPORT project and improving care for seriously ill patients. *JAMA: the journal of the American Medical Association*, *275*(16), 1228–1229; author reply 1230–1231.

Garfield E. (2006). The history and meaning of the journal impact factor. *JAMA*, *295*(1), 90–93. doi:10.1001/jama.295.1.90

Garfield, E. (1963). Science Citation Index. *Science Citation Index 1961*, *1*, v – vxi.

Gherghina, S., & Katsanidou, A. (2013). Data Availability in Political Science Journals. *European Political Science*. doi:10.1057/eps.2013.8

Gleit, C., & Graham, B. (1989). Secondary data analysis: a valuable resource. *Nursing research*, *38*(6), 380–381. doi:10.1097/00006199-198911000-00018

Google. (n.d.). About Google Scholar. Retrieved March 8, 2013, from http://scholar.google.com/intl/en-US/scholar/about.html

Grella, C. E., Hser, Y.-I., & Hsieh, S.-C. (2003). Predictors of drug treatment re-entry following relapse to cocaine use in DATOS. *Journal of Substance Abuse Treatment*, *25*(3), 145–154. doi:10.1016/S0740-5472(03)00128-4

Gutmann, M., Schurer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of social science data. *Data Science Journal*, *3*, 209–221. doi:10.2481/dsj.3.209

Harrell, F. E., Jr, Lee, K. L., Califf, R. M., Pryor, D. B., & Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, *3*(2), 143–152. doi:10.1002/sim.4780030207

Harzing, A.-W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? *Scientometrics*, *94*(1), 23–34. doi:10.1007/s11192-012-0738-1

Hedstrom, M., & Niu, J. (2008). Incentives for Data Producers to Create "Archive-Ready" Data: Implications for Archives and Records Management. Presented at the Society of American Archivists Research Forum, San Francisco, CA.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, *21*(16), 2409–2419. doi:10.1002/sim.1047

Henige, D. P. (2006). Discouraging Verification: Citation Practices across the Disciplines. *Journal of Scholarly Publishing*, *37*(2), 99–118.

Herndon, T., Ash, M., & Pollin, R. (2013). *Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff*. Retrieved from http://www.peri.umass.edu/236/hash/31e2ff374b6377b2ddec04deaa6388b1/publication/566/

Hicks, D. (2005). The Four Literatures of Social Science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 473–496). Springer Netherlands. Retrieved from http://link.springer.com.proxy.lib.umich.edu/chapter/10.1007/1-4020-2755-9_22

Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data Access, Ownership, and Control Toward Empirical Studies of Access Practices. *Science Communication*, *15*(4), 355–372. doi:10.1177/107554709401500401

Hinds, P. S., Vogel, R. J., & Clarke-Steffen, L. (1997). The Possibilities and Pitfalls of Doing a Secondary Analysis of a Qualitative Data Set. *Qualitative Health Research*, *7*(3), 408–424. doi:10.1177/104973239700700306

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. doi:10.1073/pnas.0507655102

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, *104*(49), 19193–19198. doi:10.1073/pnas.0707962104

ISI Coverage by Discipline. (2005). In *Citation Analysis in Research Evaluation* (pp. 119–136). Springer Netherlands. Retrieved from http://link.springer.com.proxy.lib.umich.edu/chapter/10.1007/1-4020-3714-7_8

Kapeller, J. (2010). Citation Metrics: Serious Drawbacks, Perverse Incentives, and Strategic Options for Heterodox Economics. *American Journal of Economics and Sociology*, *69*(5), 1376–1408. doi:10.1111/j.1536-7150.2010.00750.x

Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics — re-shaping scientific practice. *Nature Reviews Genetics*, *10*(5), 331–335. doi:10.1038/nrg2573

Kelly, C. D., & Jennions, M. D. (2006). The h index and career assessment by numbers. *Trends in Ecology & Evolution*, *21*(4), 167–170. doi:10.1016/j.tree.2006.01.005

King, G. (1995). Replication, Replication. *PS: Political Science and Politics*, *28*(3), 444–452. doi:10.2307/420301

King, G. (2003). The Future of Replication. *International Studies Perspectives*, *4*, 443–499.

King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. *Science*, *331*(6018), 719–721. doi:10.1126/science.1197872

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(2), 137–163. doi:10.2307/25791637

Lasda Bergman, E. M. (2012). Finding Citations to Social Work Literature: The Relative Benefits of Using Web of Science, Scopus, or Google Scholar. *The Journal of Academic Librarianship*, *38*(6), 370–379. doi:10.1016/j.acalib.2012.08.002

Latour, B., & Woolgar, S. (1986). *Laboratory life: the construction of scientific facts*. Princeton, N.J.: Princeton University Press.

Lavis, J., Ross, S., McLeod, C., & Gildiner, A. (2003). Measuring the impact of health research. *Journal of Health Services Research & Policy*, *8*(3), 165–170. doi:10.1258/135581903322029520

Lawrence, B., Jones, C., Matthews, 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. doi:10.2218/ijdc.v6i2.205

Lawrence, B., Lowry, R., Miller, P., Snaith, H., & Woolf, A. (2009). Information in environmental data grids. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1890), 1003 –1014. doi:10.1098/rsta.2008.0237

Leberg, P. L., & Neigel, J. E. (1999). Enhancing the Retrievability of Population Genetic Survey Data? An Assessment of Animal Mitochondrial DNA Studies. *Evolution*, *53*(6), 1961. doi:10.2307/2640454

Leydesdorff, L. (2009). Mapping Interdisciplinarity at the Interfaces between the Science Citation Index and the Social Science Citation Index. *arXiv:0911.2723*. Retrieved from http://arxiv.org/abs/0911.2723

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, *5*(1), 87–100. doi:10.1016/j.joi.2010.09.002

Line, M. B. (1979). The influence of type of sources used on the results of citation analysis. *Journal of documentation*, *35*(4).

López-Illescas, C., de Moya-Anegón, F., & Moed, H. F. (2008). Coverage and citation impact of oncological journals in the Web of Science and Scopus. *Journal of Informetrics*, *2*(4), 304–316. doi:10.1016/j.joi.2008.08.001

Lord, P., Macdonald, A., Sinnott, R., Ecklund, D., Westhead, M., & Jones, A. (2005). *Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models, The Joint Data Standards Study, http://www.mrc.ac.uk/pdfjdss_final_report.pdf* (No. UKeS-2006-02). Retrieved from http://www.nesc.ac.uk/technical_papers/UKeS-2006-02.pdf

Louis, K., Jones, L., & Campbell, E. (2002). Sharing in Science. *American Scientist*, *90*(4), 304. doi:10.1511/2002.4.304

Markson, L. J., & Clark, J. (1996). What SUPPORT really means. Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments. *Hospital practice (1995)*, *31*(9), 11–13.

Mathys, T., & Kamel Boulos, M. (2011). Geospatial resources for supporting data standards, guidance and best practice in health informatics. *BMC Research Notes*, *4*(1), 19. doi:10.1186/1756-0500-4-19

McCarty, C., Jawitz, J. W., Hopkins, A., & Goldman, A. (2012). Predicting author h-index using characteristics of the co-author network. *Scientometrics*, 1–17. doi:10.1007/s11192-012-0933-0

McDade, L. A., Maddison, D. R., Guralnick, R., Piwowar, H. A., Jameson, M. L., Helgen, K. M., … Vis, M. L. (2011). Biology Needs a Modern Assessment System for Professional Productivity. *BioScience*, *61*(8), 619–625. doi:10.1525/bio.2011.61.8.8

McDonald, M. P., & Popkin, S. L. (2001). The Myth of the Vanishing Voter. *The American Political Science Review*, *95*(4), 963–974. doi:10.2307/3117725

Meho, L. I., & Rogers, Y. (2008). Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, *59*(11), 1711–1726. doi:10.1002/asi.20874

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125. doi:10.1002/asi.20677

Microsoft. (n.d.). About Microsoft Academic Search. Retrieved March 8, 2013, from http://academic.research.microsoft.com/About/Help.htm

Mikki, S. (2009). Comparing Google Scholar and ISI Web of Science for Earth Sciences. *Scientometrics*, *82*(2), 321–331. doi:10.1007/s11192-009-0038-6

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: an expanded sourcebook*. Thousand Oaks: Sage Publications. Retrieved from http://hdl.handle.net/2027/[u]: mdp.39015050352072

Mojtabai, R. (2005). Compliance with mental health and other specialty care referrals among Medicare/Medicaid dual enrollees. *Community mental health journal*, *41*(3), 339–344.

Mojtabai, R., & Graff Zivin, J. (2003). Effectiveness and Cost-effectiveness of Four Treatment Modalities for Substance Disorders: A Propensity Score Analysis. *Health Services Research*, *38*(1 Pt 1), 233–259. doi:10.1111/1475-6773.00114

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

Moskowitz, E. H., & Nelson, J. L. (1995). The Best Laid Plans. *Hastings Center Report*, *25*(6), 3–5. doi:10.2307/3527847

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., … Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & dementia: the journal of the Alzheimer's Association*, *1*(1), 55–66. doi:10.1016/j.jalz.2005.06.003

National Institutes of Health. (2003, February 26). Final NIH Statement on Sharing Research Data (NOT-OD-03-032).

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A Review. *Scientometrics*, *66*(1), 81–100. doi:10.1007/s11192-006-0007-2

Nelson, B. (2009). Data sharing: Empty archives. *Nature News*, *461*(7261), 160–163. doi:10.1038/461160a

Nemes, S., Jonasson, J. M., Genell, A., & Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, *9*(1), 56. doi:10.1186/1471-2288-9-56

Niu, J. (2009). *Perceived Documentation Quality of Social Science Data.* (Ph.D. dissertation). University of Michigan, Ann Arbor, MI. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/63871

Noor, M. A., Zimmerman, K. J., & Teeter, K. C. (2006). Data sharing: how much doesn't get submitted to GenBank? *PLoS biology*, *4*(7). doi:10.1371/journal.pbio.0040228

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, *1*(2), 161–169. doi:10.1016/j.joi.2006.12.001

Ochsner, S. A., Steffen, D. L., Stoeckert, C. J., & McKenna, N. J. (2008). Much room for improvement in deposition rates of expression microarray datasets. *Nature Methods*, *5*(12), 991–991. doi:10.1038/nmeth1208-991

Osgood, D. W., & Anderson, A. L. (2004). Unstructured Socializing and Rates of Delinquency. *Criminology*, *42*(3), 519–550. doi:10.1111/j.1745-9125.2004.tb00528.x

Parsons, M. A., Duerr, R., & Minster, J.-B. (2010). Data citation and peer review. *Eos, Transactions, American Geophysical Union*, *91*(34), 297–298. doi:10.1029/2010EO340001

Payne, D. (2004, November 23). Google Scholar Welcomed. *The Scientist*. Retrieved from http://www.the-scientist.com/?articles.view/articleNo/23178/title/Google-Scholar-welcomed/

Peritz, B. C. (1983a). A classification of citation roles for the social sciences and related fields. *Scientometrics*, *5*(5), 303–312. doi:10.1007/BF02147226

Peritz, B. C. (1983b). Are methodological papers more cited than theoretical or empirical ones? The case of sociology. *Scientometrics*, *5*(4), 211–218. doi:10.1007/BF02019738

Perry, C. M. (2008). Archiving of publicly funded research data: A survey of Canadian researchers. *Government Information Quarterly*, *25*, 133–148.

Pestana, A., Gomez, I., Fernandez, M. T., Zulueta, M. A., & Mendez, A. (1995). Scientometric evaluation of R&D activities in medium-size institutions : a case study based on the Spanish Scientific Research Council (CSIC). In M. Koening & A. Bookstein (Eds.), *The Proceedings of the Fifth International Conference of the International Society for Scientometrics and Informetrics* (pp. 425–434).

Pienta, A. M., Alter, G. C., & Lyle, J. A. (2010). The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. Presented at the The Organisation, Economics and Policy of Scientific Research, Torino, Italy. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/78307

Piwowar, H. A. (2011). Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS ONE*, *6*(7), e18657. doi:10.1371/journal.pone.0018657

Piwowar, H. A., Becich, M. J., Bilofsky, H., Crowley, R. S., & on behalf of the caBIG Data Sharing and Intellectual Capital Workspace. (2008). Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers. *PLoS Med*, *5*(9), e183. doi:10.1371/journal.pmed.0050183

Piwowar, H. A., Carlson, J. D., & Vision, T. J. (2011). Beginning to track 1000 datasets from public repositories into the published literature. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–4. doi:10.1002/meet.2011.14504801337

Piwowar, H. A., & Chapman, W. W. (2010). Public sharing of research datasets: A pilot study of associations. *Journal of Informetrics*, *4*(2), 148–156. doi:10.1016/j.joi.2009.11.010

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

Piwowar, H. A., Vision, T. J., & Whitlock, M. C. (2011). Data archiving is a good investment. *Nature*, *473*(7347), 285–285. doi:10.1038/473285a

Porter, A. L., Cohen, A. S., David Roessner, J., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, *72*, 117–147. doi:10.1007/s11192-007-1700-5

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*, 719–745. doi:10.1007/s11192-008-2197-2

Prendergast, T. J. (1996). The SUPPORT project and improving care for seriously ill patients. *JAMA: the journal of the American Medical Association*, *275*(16), 1227; author reply 1230–1231.

Pritchard, S. M. (2005). *Informatics and Knowledge Management for Faculty Research Data* (No. ERB0502). Retrieved from http://www.educause.edu/ECAR/InformaticsandKnowledgeManagem/157535

Reed Elsevier. (2004). Scopus Comes of Age. Retrieved from http://www.reedelsevier.com/index.cfm?Articleid=1075

Reisig, M. D., & Parks, R. B. (2004). Can Community Policing Help the Truly Disadvantaged? *Crime & Delinquency*, *50*(2), 139–167. doi:10.1177/0011128703253157

Reskin, B. F. (1977). Scientific Productivity and the Reward Structure of Science. *American Sociological Review*, *42*(3), 491–504. doi:10.2307/2094753

Reyes-Ortiz, C. A., Berges, I. M., Raji, M. A., Koenig, H. G., Kuo, Y.-F., & Markides, K. S. (2008). Church Attendance Mediates the Association Between Depressive Symptoms and Cognitive Functioning Among Older Mexican Americans. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *63*(5), 480–486.

Saleem, T. (2011). The Hirsch index - a play on numbers or a true appraisal of academic output? *International archives of medicine*, *4*, 25. doi:10.1186/1755-7682-4-25

Savage, C. J., & Vickers, A. J. (2009). Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLoS ONE*, *4*(9), e7078. doi:10.1371/journal.pone.0007078

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, *314*(7079), 497–497. doi:10.1136/bmj.314.7079.497

Shankar, K. (2007). Order from chaos: The poetics and pragmatics of scientific recordkeeping. *Journal of the American Society for Information Science and Technology*, *58*(10), 1457–1466. doi:10.1002/asi.20625

Sieber, P. J. E., & Trumbo, B. E. (1995). (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics*, *1*(1), 11–20. doi:10.1007/BF02628694

Soos, S., & Kampis, G. (2010). Diversity and Polarization of Research Performance: Evidence from Hungary. *arXiv:1009.5520*. Retrieved from http://arxiv.org/abs/1009.5520

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

Strotmann, A., & Zhao, D. (2010). Combining commercial citation indexes and open-access bibliographic databases to delimit highly interdisciplinary research fields for citation analysis. *Journal of Informetrics*, *4*(2), 194–200. doi:10.1016/j.joi.2009.12.001

Svider, P. F., Choudhry, Z. A., Choudhry, O. J., Baredes, S., Liu, J. K., & Eloy, J. A. (2013). The use of the h-index in academic otolaryngology. *The Laryngoscope*, *123*(1), 103–106. doi:10.1002/lary.23569

Thomson Reuters. (n.d.). Web of Science Factsheet. Retrieved March 8, 2013, from http://thomsonreuters.com/content/science/pdf/Web_of_Science_factsheet.pdf

Tolbert, C. J., & Smith, D. A. (2005). The Educative Effects of Ballot Initiatives on Voter Turnout. *American Politics Research*, *33*(2), 283–309. doi:10.1177/1532673X04271904

Tucker, J. (2009). *Motivating subjects: Data sharing in cancer research* (Dissertation). Virginia Polytechnic Institute, Falls Church, VA.

UC Curation Center and Publishing Services, California Digital Library. (2011, June 30). *Data as Publication: Data Papers and Data Journals*. Webinar. Retrieved from http://www.cdlib.org/services/uc3/docs/UC3-webinar-datapub20110630.pdf

Uhlir, P. F., Board on Research Data and Information, Policy and Global Affairs, & National Research Council. (2012). *For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=13564

Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., & Warner, S. (2004). Rethinking Scholarly Communication. *D-Lib Magazine*, *10*(9). doi:10.1045/september2004-vandesompel

Van House, N. A., Butler, M. H., & Schiff, L. R. (1998). Cooperative knowledge work and practices of trust: sharing environmental planning data sets. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work* (pp. 335–343). New York, NY, USA: ACM. doi:http://doi.acm.org.proxy.lib.umich.edu/10.1145/289444.289508

Vines, T. H., Andrew, R. L., Bock, D. G., Franklin, M. T., Gilbert, K. J., Kane, N. C., … Yeaman, S. (2013). Mandated data archiving greatly improves access to research data. *arXiv:1301.3744*. doi:10.1096/fj.12-218164

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American journal of epidemiology*, *165*(6), 710–718. doi:10.1093/aje/kwk052

Wallis, J. C., Borgman, C. L., Mayernik, M. S., Pepe, A., Ramanathan, N., & Hansen, M. (2007). Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. In L. Kovács, N. Fuhr, & C. Meghini (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 380–391). Springer Berlin

Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-74851-9_32

Walton, D. W. H. (2010). Data Citation - Moving to New Norms. *Antarctic Science*, *22*(4), 333–333. doi:10.1017/S0954102010000520

Weber, N., & Chao, T. (2011, June 1). *A Multidisciplinary Analysis of Data Reuse Activities: An Initial Framework*. Presented at the IASSIST 2011, Vancouver, British Columbia, Canada. Retrieved from http://www.iassistdata.org/conferences/2011/presentation/2847

Whyte, A., & Pryor, G. (2011). Open Science in Practice: Researcher Perspectives and Participation. *International Journal of Digital Curation*, *6*(1), 199–213. doi:10.2218/ijdc.v6i1.182

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *The American psychologist*, *61*(7), 726–728. doi:10.1037/0003-066X.61.7.726

Wilkinson, M. (2011). Datacite: The International Data Citation Initiative: Datasets Programme. *SSRN Working Paper Series*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1743431#

Yakel, E., Faniel, I., Kriesberg, A., & Yoon, A. (2013). Trust in Digital Repositories. *International Journal of Digital Curation*, *8*(1), 143–156. doi:10.2218/ijdc.v8i1.251

Zimmerman, A. S. (2008). New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. *Science, Technology & Human Values*, *33*(5), 631 –652. doi:10.1177/0162243907306704