PERCEIVED DOCUMENTATION QUALITY OF SOCIAL SCIENCE DATA

by

Jinfang Niu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in The University of Michigan
2009

Doctoral committee:

Associate Professor Margaret L. Hedstrom, Chair
Professor Paul N. Courant
Professor Myron P. Gutmann
Assistant Professor Ixchel M. Faniel

To Grandma

ACKNOWLEDGEMENTS

This dissertation could not have happened without tremendous support from many individuals. First, I would like to thank my advisor, Professor Margaret L. Hedstrom. My greatest achievement through the Ph.D. program is the change from a narrow-minded librarian whose thoughts were dominated by intuition and emotion to a researcher with analytical and strategic thinking skills. This change was not possible without the support from my advisor. Instead of insisting that I stay in the traditional archive field, Margaret allowed me to swim in theories and methodologies from other fields by taking courses about mathematic modeling, rational choice and game theory. Also very importantly, she warned me not to stray too far away and told me when it was time to start a dissertation. She showed great patience and tolerance in reading my various drafts, gave insightful instructions and guided me step by step toward the end of this journey. Her generous funding support freed me from a teaching load and made it possible for me to focus on research. Professor Myron P. Gutmann, as the director of the Interuniversity Consortium for Political and Social Research (ICPSR) and a professor with teaching load, was willing to meet me several times to help me develop my dissertation proposal and gave very detailed instructions in presenting the results from this study. As an expert in the data curation field, his comments and suggestions were tremendously helpful, especially in thinking about why I needed to do this research, and how useful it is for the data curation field. Professor Paul N. Courant was very insightful in pointing out the most critical problem in my research. He guided me to think about how my research could be useful for policy makers, and how to extend my research to the secondary use of self-collected data. Professor Ixchel Faniel provided very detailed instructions in creating measurements for some constructs and in organizing and structuring the dissertation. Her comments on the first draft of this dissertation were very constructive, covering both big picture problems and very small details. Professor Ann Zimmerman was very helpful to

me, not only because her dissertation and publications are part of the foundation of my research, but also because she was always willing to chat with me about my ideas and give suggestions.

I would like to thank the funders of this research. They are the NSF Award # IIS 0456022, Rackham Research Grant for Dissertation and One-term Dissertation Writing Award. Thanks very much for the School of Information doctoral committee for nominating me for the One-term Dissertation Writing Award. Without this award, I would not have graduated until one semester later.

I am very grateful to numerous interviewees and survey respondents, especially interviewees who spent about one hour answering questions with great patience. I thank statisticians at the Center for Statistical Consultation and Research (CSCAR) for their assistance in data analysis. I thank data archivists from ICPSR for their support in conducting related studies that motivated my dissertation research. I also wish to thank the staff at the School of Information, particularly Sue Schuon, Jocelyn Webber and Christine Eccleston for their long-time administrative support.

I am very grateful to Beth St. Jean, Ricardo L. Punzalan, Kim Yong-mi, Maria Souden, Morgan Daniels and Xingxing Yao for their instructions in English writing, comments on my paper drafts, and help with developing my presentation skills.

I thank my friends and numerous nice people I met at Ann Arbor. They brought me an enjoyable time and taught me important knowledge about life. I greatly appreciate the sacrifice my parents made for my education. I thank my parents-in-law for their love and care. Finally and most importantly, I was quite fortunate to have a husband, Linhong Chen, who is also my best friend, a private tutor and advisor. He encouraged me to start the Ph.D. study, pulled me through the most difficult time during the first 3-4 years, and took care of every aspect of my life.

TABLE OF CONTENTS

DEDICATION..ii

ACKNOWLEDGEMENTS..iii

LIST OF FIGURES..vii

LIST OF TABLES..viii

LIST OF APPENDICES...xi

ABSTRACT..xii

CHAPTER ONE INTRODUCTION...1

Data, Social Science Data and Secondary Data Analysis……...............................3

Documentation...5

Documentation is important for secondary data analysis......................................7

Documentation is often inadequate 8

Summary...10

CHAPTER TWO LITERATURE REVIEW...11

Knowledge Senders…………………………….....................................12

Intermediaries..18

Knowledge Transferred..19

Knowledge Receivers………………………………………………..…….22

Knowledge Transfer Channels……………………………………………..23

Research Questions…………………………………………………………25

Summary...25

CHAPTER THREE METHODOLOGY...27

Exploratory Interviews 28

Survey Instrument Development...30

Target Population and Sampling Strategies 37

Pilot Survey..38

Formal Survey..39

Summary...40

CHAPTER FOUR FINDINGS AND DISCUSSIONS...41

Sample Characteristics...41

The Reliability And Validity Of DEM..46

Impacting Factors Of Perceived Documentation Quality......................................51

The Effect Of Perceived Documentation Quality On Users' Incentive To Use Secondary Data..71

How Do Users Overcome Inadequate Documentation?..75

Summary...82

CHAPTER FIVE CONCLUSIONS...84

Implications For Strategies To Help Secondary Data Use....................................85

Limitations...86

Future Work 87

APPENDICES..89

BIBLIOGRAPHY 111

LIST OF FIGURES

Figure 1 The process of secondary data analysis..8

Figure 2 General knowledge transfer..12

Figure 3 Characteristics of data…………………………………………………………….36

Figure 4 Skewed distribution of the use of datasets...44

LIST OF TABLES

Table 1 The 19 elements for completeness…………………….................................32

Table 2 Users' professional status……………………..42

Table 3 Types of data producers...44

Table 4 Most frequently used data……………………………………………….....45

Table 5 Value distributions of the four items for ease-of-use...46

Table 6 Correlations between the four sub-measures of ease-of-use…...........................47

Table 7 Value distributions of the three items for sufficiency...48

Table 8 Correlations between the four items of sufficiency..48

Table 9 Correlations between the three measures for perceived documentation quality..49

Table 10 Difference in overall quality between data produced for self-use and data produced for sharing……………………………………………………………..52

Table 11. Difference in ease-of-use between data produced for self-use and data produced for sharing……………………………………………………………………..52

Table 12. Difference in sufficiency between data produced for self-use and data produced for sharing…………………………………………………………………..…53

Table 13. Difference in ease-of-use for data produced for sharing……………………..….57

Table 14. Difference in sufficiency for data produced for sharing……………………..…..57

Table 15. Difference in overall quality for data produced for sharing…………………..58

Table 16. Difference in overall quality for data produced for self-use………………..…58

Table 17. Difference in sufficiency between qualitative and quantitative data………....60

Table 18. The difference in ease-of-use between qualitative and quantitative data……..60

Table 19. Difference in overall quality between survey data and data collected by other methods………………………………………………………………………..…61

Table 20. Difference in ease-of-use between survey data and data collected by other methods……………………………………………………………………………..61

Table 21. Difference in sufficiency between survey data and data collected by other methods……………………………………………………………………………..61

Table 22. Difference in perceived documentation quality across professional ranks…...66

Table 23. Partial correlations between perceived documentation quality and the five measurement items for absorptive capacity……………………………………...67

Table 24. Partial correlations between perceived documentation quality and absorptive capacity……………………………………………………………………………..68

Table 25. Sources for outside information………………………………………………79

Table 26. Reasons for obtaining outside information………………………………...…80

Table 27. Combined reasons for obtaining outside information…………………………81

Table 28. Value distributions of the 19 items for completeness……………………...…90

Table 29. Means and standard deviations of the 19 items for completeness……………91

Table 30. Correlation matrix for the items of absorptive capacity………………………92

Table 31. Value distributions of four items for absorptive capacity………………….…92

Table 32. Means and standard deviations of four items for absorptive capacity………...92

Table 33: Difference in overall quality between students and professors…………….…93

Table 34. Difference in sufficiency between students and professors…………………...93

Table 35. Difference in ease-of-use between students and professors……………….…93

Table 36. Difference in ease-of-use between users who are more and less familiar with the topics of data………………………………………………………………………94

Table 37. Difference in sufficiency between users who are more and less familiar with the topics of data………………………………………………………………………94

Table 38. Difference in overall quality between users who are more and less familiar with the topics of data………………………………………………………………………94

Table 39. Difference in ease-of-use between users experienced and users non-experienced in using the same data……………………………………………………95

Table 40. Difference in sufficiency between users experienced and users non-experienced in using the same data………………………………………………………………95

Table 41. Difference in overall quality between users experienced and users non-experienced in using the same data……………………………………………………96

Table 42. Difference in ease-of-use between users who are experienced in secondary data analysis and users who are not…………………………………………………………...96

Table 43. Difference in sufficiency between users who are experienced in secondary data analysis and users who are not……………………………………………………….......96

Table 44. Difference in overall quality between users who are experienced in secondary data analysis and users who are not……………………………………………….......97

Table 45. Difference in ease-of-use between users who are experienced in collecting and analyzing primary data and users who are not……………………………………........97

Table 46. Difference in sufficiency between users who are experienced in collecting and analyzing primary data and users who are not……………………………………........98

Table 47. Difference in overall quality between users who are experienced in collecting and analyzing primary data and users who are not………………………………………98

LIST OF APPENDICES

Appendix A. Supplementary Tables..90

Appendix B. Survey Instrument...99

Appendix C. Interview Invitation Letter..105

Appendix D. Survey Invitation Email (to secondary data users)....................................106

Appendix E. Survey Invitation Email (to data librarians)..107

Appendix F. Consent Form (Survey)...108

Appendix G. Consent form (Interviews)...110

ABSTRACT

This study is about user perceived documentation quality of social science data. The goals are to identify impacting factors of perceived documentation quality and find out how perceived documentation quality affects secondary data use. To help identify the impacting factors of perceived documentation quality, documentation of social science data was investigated as a knowledge transfer channel between data producers and secondary data users. A general knowledge transfer model was formulated based on literature in knowledge management and then applied to characterize the knowledge senders, receivers, knowledge transferred, and knowledge transfer channels in secondary data use. In doing this, four possible impacting factors of perceived documentation quality were identified: data producers' incentives and ability, data users' absorptive capacity, the existence of intermediaries between data producers and users, and data's vulnerability to the tacit knowledge problem. Hypotheses about how each factor affects perceived documentation quality were formulated.

Interviews and surveys were conducted with secondary users of social science data. Preliminary analysis of the interviews helped the survey design, such as deciding the survey population and units of analysis, and identifying indicators of users' absorptive capacity. A Documentation Evaluation Model (DEM) was constructed as a tool to assess perceived documentation quality. The reliability and validity of DEM were tested based on the survey and interview data collected for this study. I found DEM was reliable and valid in general with several exceptions. Hypotheses tests proved that effects of the proposed four impacting factors of perceived documentation quality. Data produced for sharing are better documented than data produced for self-use. Users with stronger absorptive capacity tend to perceive the documentation they use as better than users with weaker absorptive capacity. Intermediaries such as data archives have been effective in improving documentation quality of data produced for sharing. Data less vulnerable to the tacit knowledge problem, such as quantitative survey data and data about straightforward facts, are perceived as better documented than data more vulnerable to the tacit knowledge problem, such as qualitative data. Inadequate documentation increases the costs of use and may turn users away in some situations. However, users' incentives to use secondary data mostly depend on how well the data fit their information needs rather than documentation quality. A well-documented dataset will not be used if it doesn't answer users' research questions. Users will not give up using a dataset simply because it is poorly documented. Their decisions to use or not depend on how much they can benefit from using the data, the cost of overcoming inadequate documentation, and the potential cost to collect the same data. Users often need to seek information not provided in documentation because of inadequate documentation (insufficient, hard to use, inaccurate), inherent limitations of documentation, users' low absorptive capacity, convenience, and for social and psychological reasons. In seeking outside information, users tend to consult multiple sources and channels. Publications based on the same data are the most frequently used outside source of information. Email or telephone is the primary channel through which users seek outside information from people.

CHAPTER ONE

INTRODUCTION

For the past several decades, more and more data have been shared or at least required to be shared with the public. The implementation of the Freedom Of Information Act (FOIA) in 1966 has made administrative records collected by government agencies available to the public (Halstuk & Chamberlin, 2006). The 1999 revision of the Office of Management and Budget (OMB) Circular A-110, otherwise known as the “Shelby Amendment”, "require[s] federal awarding agencies to ensure that all data produced under an award will be made available to the public under the FOIA” (http://www.whitehouse.gov/OMB/fedreg/a-110rev.html). In line with this legal requirement, funding agencies and scientific journals have started to require data sharing as well, such as the National Science Foundation (NSF) (http://www.nsf.gov/pubs/gpg/nsf04_23/6.jsp), the National Institute of Justice (NIJ) (Wilson & Maxwell, 2006), and the National Institutes of Health (NIH). To publish in scientific journals, U.S. scientists involved in the field of crystallography must deposit their data in the Protein Data Bank (PDB) (Arzberger et al, 2004). The same trend is happening in other countries as well. In the United Kingdom, the Economic and Social Research Council (ESRC[1]) requires all award holders to offer their computer-readable data for deposit to a data archive within three months of the end of an award (ESRC, 2000).

It is widely acknowledged that research data sharing will save funding and avoid repeated data collecting efforts, facilitate open science and deter scientific fraud. In addition,

[1] ESRC is the UK's leading research and training agency addressing economic and social concerns.

according to Lesk (2004), data sharing is causing profound changes in research methodology, the scientific paradigm has shifted from the “old style” (hypothesize, design experiment, run experiment, analyze results, evaluate hypothesis) to the “new style” (hypothesize, look up data to test it, evaluate hypothesis). An undeniable truth is that the benefit of data sharing can only be reaped through the secondary use of data. However, there is a lack of research about secondary use of social science data. Corti & Bishop (2005) pointed out that published literature on the secondary analysis of qualitative data is sparse. Actually there is a lack of research regarding secondary analysis of quantitative data as well.

It is true that numerous books and papers have been published on the topics of data sharing and secondary data analysis. But many of them are not about social science data. They are about ecology data (Zimmmeman, 2003; Borgman, Wallis, and Enydy, 2006; Borgman, Wallis, and Enydy, 2007), earth engineering, HIV/AIDS and high energy physics data (Birnholtz and Bietz, 2000), environmental data (Van House, Butler and Schiff, 1998), geology and paleontology data (Chinn & Brewer, 2001), and medical and life science data (Campbell et al., 2002; Reidpath and Allotey, 2001; Blumenthal, et al., 2006). Publications about social science data are mostly about data producers and intermediaries such as data archives. Heavily discussed are issues related to the incentives of data producers to share data. Those issues cover costs, obstacles, benefits and risks of sharing data, and the legal environment including laws, policies and professional standards governing data sharing (Siber, 1991; National Research Council, 2003). There are many publications about the data collections, roles and services of data archives written by people affiliated with data archives (Corti and Bishop, 2005; Gutmann, et al., 2004), or by experienced researchers to guide novice users where to search and how to obtain data. There are books about secondary use experiences (Dale, Arber and Procter, 1988; Bowering, 1984; Kiecolt and Nathan, 1985) and papers (David, 1991; McCall and Applebaum, 1991) that give step-by-step instructions about how to manipulate and analyze social science data. These are more like research methodology instructions than user studies, or individual’s opinions and arguments based on personal stories or anecdotal evidences, or summaries of discussion at conferences (David, 1991). It is hard to generalize those experiences.

To enable secondary data use, data and knowledge about data needs to be transferred from data producers to secondary users. The term "documentation" is used for knowledge about data recorded and transferred to secondary users that helps secondary users understand and use data. The quality of documentation plays a critical role in data sharing and secondary analysis. However, even though it is often reported that inadequate documentation is a barrier in secondary data analysis, documentation of social science data has rarely been the focus of existing studies.

This study is trying to address the void of studies on secondary users and documentation. It will study user perceived documentation quality. The goals are to identify impacting factors of documentation quality and to find out how perceived documentation quality affects secondary data use. Findings from this study will inform decisions about how to improve documentation quality and facilitate secondary data use. In the following sections, I will give detailed definitions of social science data, secondary data analysis and documentation, and illustrate the role of documentation in secondary data analysis.

Data, Social Science Data and Secondary Data Analysis

The word "data" means different things to different people in different contexts. Some people refer to everything digital as "data". For example, " 'data' denotes whatever records are stored in a computer" (Buckland, 1991). For practical purposes, it makes more sense to define data in specific contexts rather than give a general definition. This study is about secondary data analysis for social science research. My discussions will focus on social science data. In (National Research Council, 1997, p. 198), scientific data is defined as "scientific or technical measurements, values calculated there from, and observations or facts that can be represented by numbers, tables, graphs, models, text, or symbols and that are used as a basis for reasoning or further calculation" In the Office of Management and Budget's (OMB) Circular A-110 (http://www.whitehouse.gov/omb/circulars /a110/a110.html) and in the NIH data sharing policy (National Institutes of

Health, 2006), research data is defined as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues." There is a common theme in the above two different definitions: data are used as basis for reaching conclusions. Based on this common theme, I define data as "information and materials (including both information and the media that carry information) used for reaching conclusions." Research data are situational. The same information or materials may be research data for some people but not others. Likewise, the same information may be research data for a person at one time point, but not data any more at another time point, depending on whether that person uses that information or material to reach conclusions. In other words, it is about *when* are data, but not *what* are data. Photos of a civil war hero in a historical archive are regular archives. But when used by a researcher to study the civil war, those photos become data for that researcher. A poem written 500 years ago is just a poem, but it becomes datum for a researcher who studies ancient literary style.

Boruch (1985) defined data sharing as "the voluntary provision of information from one individual or institution to another for purposes of legitimate scientific research" (p. 89). Boruch's definition fits the context of this study, but I do not agree that data sharing has to be voluntary. I define data sharing for scientific research as "the provision of data from one individual or institution to another for research purposes." Based on a survey of the uses of the term "secondary data analysis" in existing literature,[2] here I give a general definition of secondary data analysis: "the analysis of data for a different purpose than what the data were originally collected for, possibly by the original data producers themselves, or in collaboration with other people, or by entirely different people." In this study, cases in which data producers analyze their own data for a different purpose are excluded, because interesting social issues involved in sharing between people, such as the incentives to share data, the cost of documenting data for other users, the uncertainties in using data collected by other people, do not exist in those cases. In this study,

[2] I found that literature by searching the library databases of the University of Michigan using the term "secondary data analysis".

secondary data analysis means the analysis of data produced by other people. More specifically, the users of data are not involved in the production process of the data directly (collect the data themselves) or indirectly (being part of the team who originally design the study and collect the data). Secondary data are also situational. It is about *when* are secondary data, not *what* are secondary data. Police departments collect vehicle crash records routinely for administrative purposes. Those records are primary data for the police department, but secondary data when used by a criminology professor for research. The Social Security Administration collects primary data about citizen's tax and welfare. Those data become secondary when used by an economist to study social and economical factors. The terms "secondary data use" and "secondary data analysis" are used interchangeably in this study. Even though secondary data analysis could be conducted for various purposes, such as teaching or decision making, this study only looks at secondary data analysis for social science research.

Documentation

In order to enable secondary data analysis, knowledge about data needs to be transferred from data producers to secondary users. Some knowledge about data is recorded and transferred to secondary users to help them understand and use data. We call that recorded knowledge documentation. The format and content of documentation vary with different data. Here are exemplary documents that can be parts of documentation: codebooks (sometimes called data dictionary), reports about the data collection project, data collection instruments, previous publications based on the data, user guides or handbooks, statistical manuals, data extraction software, program that makes new variables based on the original data, original Institutional Review Board (IRB) materials, workflow for creating new datasets based on existing data, etc. Documentation is sometimes called metadata. The word "metadata" has similar but broader meanings. Metadata is "data about data". Metadata can be categorized on different dimensions. Based on the purposes of metadata, there are metadata for resource discovery, metadata for preservation, metadata for administration, etc. Documentation for social science data is a kind of metadata used for resource discovery (searching and judging the relevancy of the data) and secondary analysis. Based on how metadata elements are organized, there are structured and unstructured metadata. Structured metadata have a relatively stable and fixed structure. For example, a MAchine Readable Catalog (MARC) record or a Dublin Core metadata record is structured metadata about an item (book, article, web page, etc). In the context of social science data sharing, a Data Documentation Initiative (DDI) record is an example of structured metadata. Data collection instruments, such as the text of interview or survey questions are examples of unstructured metadata. According to Gutmann, et al. (2004), four types of metadata are needed for using and archiving social science data. (1) Study-level metadata, also known as abstracts, study descriptions, or metadata records. This is the highest level of metadata, describing the study or collection as a whole. These metadata outline the purpose of the study, the major conceptual categories studied, the characteristics of the sample, measures, etc. (2) File-level metadata. These describe the properties of individual files in a data collection. (3) Variable-level metadata. This type of metadata describes individual measures or groups of measures. These are recorded in documentation such as codebooks and data definition statements and are essential to effective and accurate interpretation and use of data. (4) Administrative and structural metadata. These are critical to ongoing maintenance and preservation of the electronic data collections. Documentation for secondary analysis refers to the first three categories of metadata. The word "documentation" is widely used in the social science data community.

The formats of documentation vary from penciled notes and print-outs to digital Word or PDF files, web pages, Wiki pages or emails. Users may download data and documentation from the websites of data archives or data producers, or obtain them from data archives or data producers through postal mail or email. Documentation for some very large data sets is often published online as hyperlinked web pages. For example, the documentation of the General Social Survey is at: http://www.norc.org/GSS+Website. The documentation includes the codebook for the survey, related reports and publications based on the General Social Survey.

Documentation is important for secondary data analysis

Corti (2005) pointed out that the first major step in making data fully shareable is producing rich and full documentation. Here I describe a secondary data use case to illustrate how documentation is used.

Before analyzing secondary data, the first step is to search, choose and obtain the data. Users match their research interests with the descriptions of available datasets and decide which dataset to choose. According to Dale, Arber and Procter (1988), users choose a dataset based on the following information about the data: sampling procedure, population sampled, method of data collection, response rate, characteristics of non-respondents, documentation available on the study, who conducted the study, what publications have been produced, etc. That metadata information is often extracted from documentation and put into an online catalog so users can search and do a preliminary assessment of the relevancy of the data without seeing the whole documentation package. After obtaining the data, users do standard checks on the data (Bowering, 1984). Data checks include file verification and sample verification. File verification requires the researcher to determine whether the variables specified in the documentation are present in the file, whether the summary statistics given in the documentation for these variables can be replicated from the data, and whether the file contains the number of cases or records reported in the documentation. A complementary task is to find out how non-responses and missing data are handled in the dataset. To verify the definitions of variables, the user must first review the descriptions in the documentation, then list the values for the variables of interest in the total file or in a sample of the file and see whether the values on the list fit the definitions. Sample verification processes include investigating the degree to which the data reflect the sampling procedures described in the documentation, and investigating the effects and extent to which non-response, data loss and missing data affect the data. Through the checks, users get a good understanding of the data. The final step is data manipulation and analysis. Data manipulation includes constructing new variables based on the variables existing in the data, recoding the data, merging several files into one, etc. When those steps are completed, users can apply data analysis techniques to the newly created dataset. Users' data analysis skills and capabilities dominate this final step. Data manipulation and analysis should be based on good understanding of the data which relies on good documentation; otherwise data may be misused or incorrect conclusions might be drawn from the data.

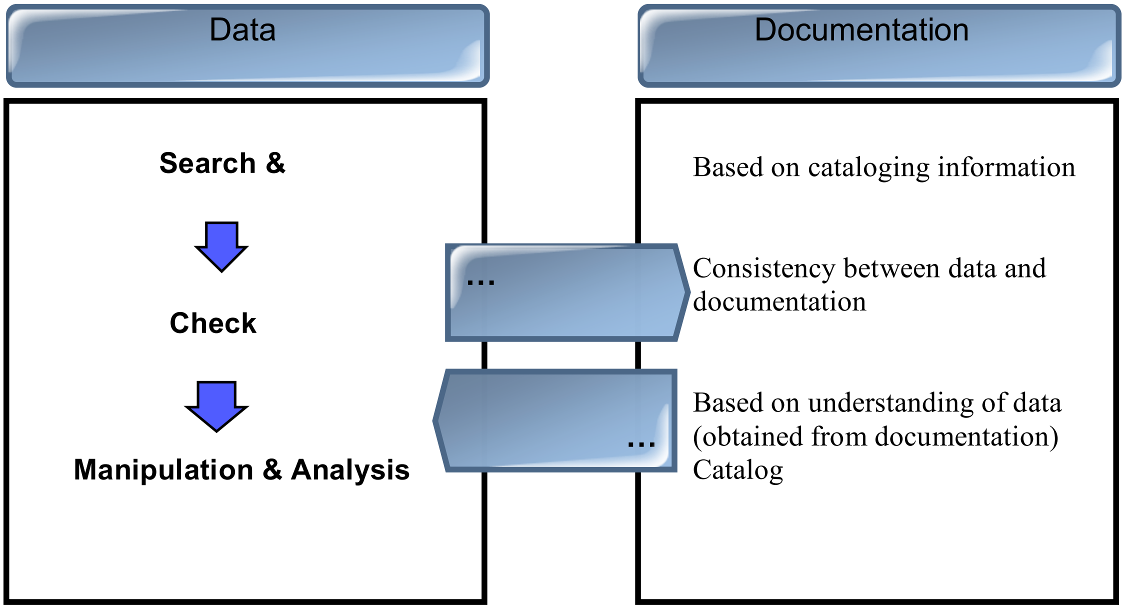
Data
Documentation
Search &
Check
Manipulation & Analysis
Based on cataloging information
...
Consistency between data and documentation
...
Based on understanding of data (obtained from documentation) Catalog

Figure 1. The process of secondary data analysis

Documentation is often inadequate

Despite its important role, documentation is commonly reported as inadequate for secondary use (Fienberg, Martin and Straf, 1985; Clubb, et al, 1985; Van Den Berg, 2005; Corti, 2000; McCall and Applebaum, 1991; Zimmerman, 2003; Borgman, 2007) (Rogers, et al, 2006). Various measures have been taken to improve the quality of documentation. Some funding agencies provide financial support for the costs involved in data sharing. In the United States, the NIH explicitly allows applicants to request funds for data sharing and archiving in their grant applications. But available financial resources were seen as inadequate for elaborate data preparation and documentation (Fienberg, Martin, & Straf, 1985). A recent survey of National Institute of Justice (NIJ) grantees found that many of them expected more financial support for documentation (Niu & Hedstrom, 2007a). Arzberger, et al. (2004) pointed out that scant attention is paid to data management in many areas of public research. NIH data sharers are allowed to charge data requestors for the costs associated with sharing data (National Institutes of Health, 2003). There is no empirical study that shows how effective this strategy is for improving documentation quality.

Data archives have tried to provide instructions to data producers about preparing documentation and about reducing the cost of this process. The social science data archive community has created a standard specifically for documenting social science data -- Data Documentation Initiative (DDI) (ICPSR, 2005). This standard provides a comprehensive list of items and elements that should be provided along with data in order to facilitate secondary use. Documentation that compiled with DDI is expected to help data producers to prepare data for secondary use. Unfortunately, a 2006 survey of NIJ grantees (who are required to deposit their data into a data archive) found that 68% of them were unfamiliar with DDI. Of those who are familiar with DDI, 29% don't know how to use it and 12% think that it further complicates the documentation process. Only 17% gave a positive response regarding DDI, indicating that it reminds them of the documentation items they need to provide (Niu & Hedstrom, 2007a). Some data archives, such as the Inter-University Consortium For Political And Social Research (ICPSR), provide detailed guidelines about how to prepare data for deposit. Results of the aforementioned 2006 survey show that the data documentation guidelines provided by ICPSR are considered to be useful; however, many depositors are not aware of them (Niu & Hedstrom, 2007a). Data archives have also proposed the idea of early documentation, which entails fleshing out a data sharing and archiving plan while the researcher is still at the stage of outlining and writing the grant application (Jacobs & Humphrey, 2004; ICPSR, 2005). Early attention to documentation is expected to reduce the cost of documentation. However, it is unclear how the early documentation concept has been implemented in practice. The 2006 survey mentioned earlier also found that data producers still complain about the problems caused by documenting at the end of research projects (Niu & Hedstrom, 2007a).

Summary

For the past several decades, more and more data have been shared or at least are required to be shared with the public. It is widely acknowledged that research data sharing will save funding and avoid repeated data collecting effort, facilitate open science and deter scientific fraud. The benefit of data sharing can only be reaped through the secondary use of data. However, there is a lack of research about secondary use of social science data. To enable secondary data use, data and knowledge about data need to be transferred from data producers to secondary users. Documentation is metadata of data. It is an important knowledge transfer channel between data producers and secondary users. However, even though it is often reported that inadequate documentation is a barrier in secondary data analysis, documentation of social science data has rarely been the focus of existing studies. This study is trying to address the void of studies on secondary users and documentation. It will study user perceived documentation quality. The goal is to identify impacting factors of documentation quality, and find out how perceived documentation quality affects secondary data use. Findings from this study will inform decisions about how to improve documentation quality and facilitate secondary data use.

CHAPTER TWO

LITERATURE REVIEW

To enable secondary data analysis, data and knowledge about data needs to be transferred from data producers to secondary users. To better understand this knowledge transfer process, a theoretical guide is needed. After a thorough search of existing literature, I did not find a general theoretical framework that fits perfectly. However, Szulanski's (1996) model of the difficulty of knowledge transfer within firms contains some useful elements. According to Szulanski, the four influential factors for the difficulty of knowledge transfer are: the knowledge transferred, the source (sender), the recipient (receiver) and the context in which the knowledge transfer takes place. Based on the model in Szulanski (1996), concepts and ideas in other literature, and my own thoughts, I formulated a general knowledge transfer model: knowledge transfer is the process of transferring knowledge from the sender to the receiver through some channel. Knowledge transferred may be explicit knowledge only, or include a tacit component. To transfer knowledge successfully, the sender should be willing to, and capable of, transferring the knowledge. The receiver should have an incentive to accept the knowledge and the ability to absorb the knowledge. Effective channels need to be available for this knowledge transfer. In this chapter, the model is used as a framework to integrate existing research about data sharing and secondary data analysis. I will characterize the knowledge senders, receivers, knowledge transferred and the transfer channels in secondary data use. In doing this, the impacting factors of documentation quality will be identified, and hypotheses about how each factor affects perceived documentation quality will be formulated.

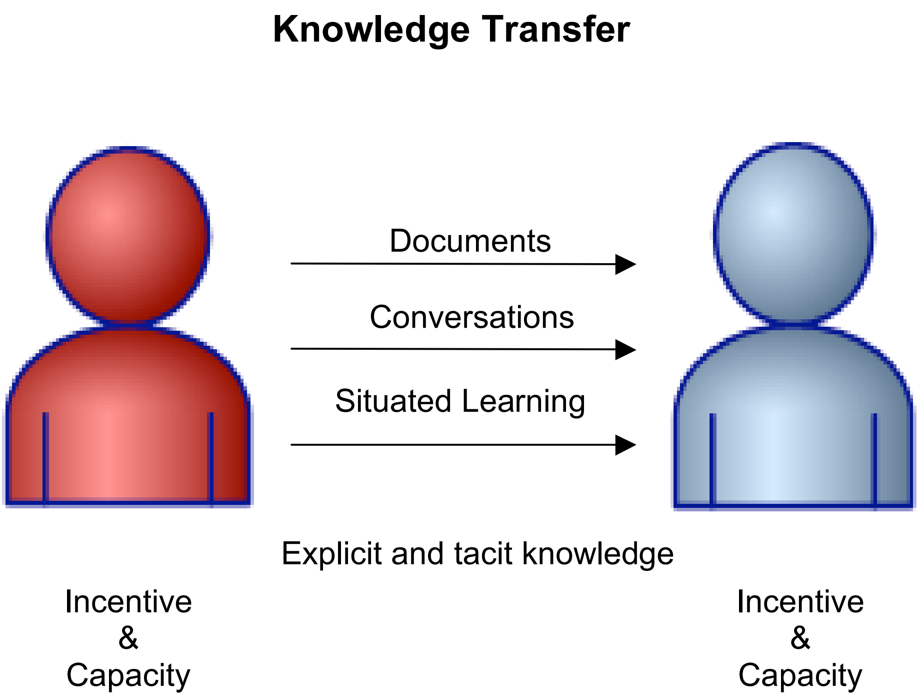

Figure 2. General knowledge transfer

Knowledge Senders

In secondary data use, the knowledge senders are data producers. According to the DDI international standard (http://www.ddialliance.org/DDI/dtd/version2-1-all.html), data authors are the person(s), corporate body(ies), or agency(ies) responsible for the work's substantive and intellectual content. Data authors are also called primary investigators in DDI. Data producers are the person(s) or organization(s) with the financial or administrative responsibility for the physical processes whereby the document was brought into existence. Data producers and authors can be one entity or two separate entities. In this dissertation, when I say data producer, I mean both data authors and data producers defined in DDI.

Incentives

Some data are produced with the intent to be shared with many people. Those data are typically produced by survey research organizations, commercial data companies who sell data for profit, and some government agencies that have a tradition of producing and sharing data, such as the Census Bureau, Bureau of Labor Statistics, and the Center for Disease Control. The mission of those producers is to collect high quality data for other researchers. Their performance is evaluated by how many users they attract and how much valuable research is conducted using their data. Those data producers are highly motivated to share data, document data well, and provide user assistance. The incentive to share and document data is not an issue for this type of data.

When data are produced for self-use, the incentive to share is much more complicated. Existing literature suggests that researchers' willingness to share scientific data is conditioned on various factors. Researchers may be willing to share their data after publication under certain conditions (e.g., attribution required, non-commercial use only) (Borgman, Wallis, & Enyedy, 2007). Such researchers are generally more willing to share contextual data than experimental data[1] (Borgman, Wallis, & Enyedy, 2007). Researchers' willingness to share increases with the degree of automation and decreases with the amount of effort expended for data production (Pritchard, Carver, & Anand, 2004; Borgman, Wallis, & Enyedy, 2007). Scholars whose data collection and analysis mostly automated are the most likely to share their raw data and analyses. Those whose data collection and analysis least automated and most labor-intensive are the most likely to guard their data. Data sharing is more common in big science than in small science[2] fields (Estrin, Michener, & Bonito, 2003; Zimmerman, 2003). Those most likely to use secondary data were the most likely to share (Zimmerman, 2003; Borgman, Wallis, & Enyedy, 2007). Researchers may wish to share their data if they do not have enough time or capability to analyze the data well, or when the data are much larger than necessary to study the phenomenon specifically being investigated by the data producers (Birnholtz & Bietz, 2000). The reality of sharing social science data produced for self-use is

[1] Experimental data are those that reflect the hypotheses and research questions of the investigator. Contextual data include micrometeorological measurements (temperature, humidity, etc.) and calibration of tools and instruments for the study, such as the density of shade cloth for a field experiment.

[2] Big Science reflects the large, complex scientific endeavors in which society makes major investments. These are characterized by expensive equipment that must be shared among many collaborators, such as particle accelerators or space stations. Little science is the independent, smaller scale work.

disappointing according to Wicherts, et al. (2006). They originally aimed to reanalyze datasets reported in 141 empirical articles recently published by the American Psychological Association (APA). In June 2005, they contacted the corresponding author of every article that appeared in the last two 2004 issues of four major APA journals. Because those articles had been published in APA journals, they were certain that all of the authors had signed the APA Certification of Compliance With APA Ethical Principles, which includes the principle on sharing data for reanalysis. Unfortunately, six months later, after writing more than 400 e-mails, and sending some corresponding authors detailed descriptions of their study aims, approvals of their ethical committee, signed assurances not to share data with others, and even their full resumes, they ended up with a meager 38 positive reactions and the actual datasets from 64 studies (25.7% of the total number of 249 data sets). This means that 73% of the authors did not share their data.

Disincentives to share may arise due to various reasons. Few institutions have formal policies and procedures governing access to, and retention of, research data (Council on Governmental Relations, 2006). Even though some funding agencies do have data sharing policies, there may be a lack of enforcement or credible threat for non-compliance. Reviewers of subsequent grant proposals assess the published products of previous research, yet they rarely put much weight on whether the data produced with prior grants were made available for use by others (Council on Governmental Relations, 2006.) Ownership of data is often ambiguous or unknown to data producers (Borgman, 2007, p. 174). Because of the complexity of intellectual property laws, some researchers (especially those in the social sciences and humanities) do not know what rights they have over their data. In the United States, OMB Circular A-110, NIH Grants Policy, and the National Science Foundation all give institutional grantees the right to own the data produced by investigators in their institutions (Council on Governmental Relations, 2006). Scholars rarely possess legal ownership of the data they produce; however, they typically feel a sense of ownership (Borgman, 2007, p. 174).

Another disincentive to data sharing is that data producers may fear that they will be harmed as a result. When data is released to the public, data producers lose the exclusive right to publish follow-up papers based on the data. Additionally, the data may no longer be offered in exchange for others' data or for funds, equipment, and other resources (Borgman, 2007, p. 174). Some data producers worry about the potential for misuse and misinterpretation of the data by unqualified users, and about the possibility of being charged with misconduct. At the same time, some data sharing policies preclude researchers from reaping the benefits they expect to derive from sharing data. For example, the National Institutes of Health (2003) and the National Research Council (2003) do not allow the investigators who produced the data to require co-authorship or collaboration as a condition of providing the data.

The cost of preparing data for sharing is another disincentive. In order to make data usable by secondary users, data producers often need to prepare their data. Data preparation entails both the compilation of the data and the creation of accompanying documentation. This process includes checking the integrity and consistency of data, careful naming of variables and selection of variable labels, organizing the variables, such as grouping them to enable secondary analysts to quickly get an overview of the data, etc. When human subjects are involved, anonymization of the data before sharing is critical. All direct identifiers of individuals, such as names, addresses, telephone numbers, and Social Security Numbers need to be removed (unless they are necessary for data analysis), indirect identifiers[3] and other information that could lead to "deductive disclosure" of participants' identities should also be removed or processed in such a way that the usefulness of the data for secondary analysis is not impacted (ICPSR, 2005). Due to the cost of data preparation, resource constraints can be a factor that keeps data producers from creating good documentation. Most respondents of our 2006 survey[4]

[3] Indirect identifiers are variables that will not, in and of themselves, reveal the identity of human subjects, but that may do so when used in combination with other variables. For example, the inclusion of a ZIP code may not be troublesome by itself, but when combined with other attributes like race and annual income, a ZIP code may allow unique individuals (i.e., extremely wealthy or very poor) residents of that ZIP code to become identifiable.

[4] This survey was conducted on grantees of the National Institute of Justice, who are required to deposit their data to a data archive at the end of their grants.

expected that more time and financial support from funding agencies could be provided for documenting data. Some data producers complained that their grant was even insufficient to cover research needs, let alone creating documentation (Niu & Hedstrom, 2007a). By the time that the data is complete and the reports delivered, time and funding are usually used up, and nothing is left for documenting and sharing data.

It could be argued that data producers need to do some of this data preparation work for their own research anyway; however, data preparation sufficient for self-use is very different from that necessary for secondary users. First, data producers do not need to process their data in order to protect the privacy of their human subjects if they retain exclusive access to their data. Second, documenting data sufficiently for others requires considerably more time and effort than documenting it for use by a small research team. Documenting data for later use also requires much more effort than what is required in order to publish data summaries in a journal article or conference paper (Borgman, Wallis, & Enyedy, 2007). Documentation for self-use tends to be incomplete and informal. The initial investigators often can accomplish their research goals without carefully cleaning and documenting their data. Some researchers keep details of their data collection and variable construction processes and the particular quirks of the data in their memories and do not put them in writing (Fienberg, Martin, & Straf, 1985; Breusch & Holloway, 2004). In other words, documenting data for secondary use requires data producers to make much tacit knowledge about their data explicit. There is cost associated with articulating tacit knowledge. According to Nelson & Winter (1982, p. 82), "Whether a particular bit of knowledge is in principle articulable or necessarily tacit is not the relevant question in most behavioral situations. Rather, the question is whether the costs associated with the obstacles to articulation are sufficiently high so that the knowledge in fact remains tacit." Data collectors sometimes prefer data preparation and documentation practices with which they are familiar, although these practices may be at odds with accepted standards (Fienberg, Martin, & Straf, 1985). Documents intended solely for self-use do not have to be technically accurate and correct (Orlikowski, 1995), and often contain shorthand detail that makes sense only to the author (Markus, 2001). Documentation for self-use tends to be useful to the author only in the short term.

Eventually some documents become difficult for the producers themselves to use, let alone secondary users (Moran, et al., 1996). When people have purposely created records for themselves, they may strenuously resist making these records public, or they may need to take extra effort to re-shape the documentation for secondary use. The effort required to explain one's data adequately increases as a function of the knowledge distance between data producers and users. Documenting research data for use by team members is more difficult than documenting it for personal use. Documenting it for off-site collaborators is more difficult still. Most difficult of all, however, is documenting for unknown future users (Borgman, 2007, p. 167; Markus, 2001). This is precisely the case with public data sharing.

In summary, when data are produced for self-use, data producers often have no incentive to share their data. Even if they are required to share their data according to FOIA or data sharing policies of funding agencies, due to the cost of preparing data for sharing, they may "comply with the letter of the law rather than its spirit, depositing poorly documented data that is of little value." (Borgman, 2007, p242).

The ability to transfer knowledge

Explicit knowledge is generally easier to transfer than tacit knowledge. When the knowledge to be transferred includes a tacit component, it is often necessary to transform the tacit knowledge into explicit knowledge before it can be transferred. The sender's ability to articulate tacit knowledge determines the potential success of the knowledge transfer. Difficulties in articulating tacit knowledge may be caused by an unawareness of the receiver's needs. When the sender has inaccurate assumptions about the receiver or when the sender and receiver have a mismatched focus, some knowledge may fail to be articulated (Collins, 2001). The difficulties in articulating tacit knowledge are likely to happen to data producers as well. Since it is hard to anticipate the needs of secondary users, the difference between documentation for self-use and secondary use challenges data producers' ability to document data well for secondary use.

People vary in their ability to articulate information. A good instructor can articulate for the students much of the knowledge that ordinarily remains tacit (Nelson and Winter, 1982, p. 78). In secondary data use, producers who produce data for sharing tend to be professional data producers, such as the aforementioned survey research organizations and data companies. They are likely to be more capable of documenting data well for secondary use.

To sum up, documentation quality is likely to be affected by the incentives and ability of data producers to document data. **Hypothesis 1: Data produced for sharing are perceived as better documented than data produced for self-use.**

Corti (2000) mentioned that the size of data producers affects their incentive to document data. When the data producer is a large group or organization, it is often the case that different individuals are involved in designing the study, collecting the data, processing the data, and analyzing the data. Thus, in order to enable other people in the producer group to understand the data, each team member has to make explicit the ways he proceeds. The larger the data producer group, the more loosely coupled the group members, the more reliant they are on documentation to transfer knowledge even among the producer group members. Corti (2000) argued that this provides an incentive to create high quality documentation. In cases where a single researcher collects and interprets data, many assumptions, procedures, processes, and decisions tend to remain undocumented tacit knowledge (Carlson and Anderson, 2007). In addition, such researchers often lack incentives to document the data for secondary use. **Hypothesis 2: Data produced by large producers are perceived as better documented than data produced by single researchers or small research groups.**

Intermediaries

Data are distributed to users either directly by data producers or through intermediaries such as data archives. Some users download data or buy CDs directly from the website of data producers. Some data producers allow archives to process and distribute their data, such as the producer of National Comorbidity Survey Replication (NCS-R). Funding agencies like National Institute of Justice (NIJ) require their grantees to deposit data into a designated data archive. Intermediaries process and enhance the quality of data and documentation before distribution. In other words, they reduce the cost to data producers and supplement the ability of data producers to document data. Therefore intermediaries are likely to have an impact on documentation quality. **Hypotheses 3: Data distributed by intermediaries are perceived as better documented by data distributed by data producers.**

Knowledge Transferred

Polanyi (1962) classified human knowledge into two categories: explicit and tacit. Explicit knowledge refers to knowledge that is transmittable in formal, systematic language. Tacit knowledge, on the other hand, is hard to formalize and communicate as it is deeply rooted in action, commitment, and involvement in a specific context (Nonaka, (1994). In Polanyi's words, it "indwells" in a comprehensive cognizance of the human mind and body. From my point of view, Polanyi and his followers differentiate tacit knowledge and explicit knowledge based on whether it is technically difficult to articulate. If it is easy to articulate, it is explicit knowledge; otherwise, it is tacit knowledge. Another type of tacit knowledge may not be technically hard to articulate, but is sensitive and subtle. Even though people may know implicitly, it is socially inappropriate to make certain knowledge explicit. In talking about why groupware fails, Grudin (1994) said a priority-based meeting scheduler failed because participants were reluctant to acknowledge publicly that some of their meetings have low priority. Other scholars differentiate tacit and explicit knowledge simply based on whether or not the knowledge has been articulated. If it has been articulated, it is explicit knowledge; otherwise, it is tacit knowledge. For example:

> Collins (2001) defines tacit knowledge as "knowledge or abilities that can be passed between scientists by personal contact but cannot be, or **have not been**, set out or passed on in formulae, diagrams, or verbal descriptions and instructions for action." (p. 72)

> Markus (2001) states that explicit knowledge is "knowledge that has at minimum been

> ‘captured’ and articulated and has ideally been ‘codified’, that is, documented, structured and disseminated” and that tacit knowledge is “knowledge that resides in people’s heads or ‘muscle memory’ and may be destined to remain there” (p. 58).

These definitions define tacit knowledge more broadly, including not only knowledge that is hard to articulate, but also knowledge that has remained tacit simply because it has not been made explicit. According to this type of definition, knowledge that is easy to articulate, but has not been articulated for some reason, is also tacit knowledge. Some researchers use the term “implicit knowledge” for knowledge that is expressible but that has not yet been made explicit (Wilson, 2002). According to the theory of communication reductionism, not everything can, or should be transferred during communication. Some kind of reduction, and thus loss of complexity is inevitable (Strathern, 2005; Carlson and Anderson, 2007). Therefore tacit knowledge defined in the broad sense is often unavoidable in human communication. In secondary data use, documentation contains explicit knowledge about the data. Insufficient documentation means that some knowledge is necessary for secondary use but not provided with the documentation. Various kinds of tacit knowledge could be included in that missing knowledge, such as knowledge that is technically hard to articulate or socially sensitive to articulate.

Some situations are more vulnerable to tacit knowledge problem thans others. In other words, verbalized or documented explicit knowledge can be sufficient in some situations, whereas in other situations, the tacit knowledge is indispensable and has to be transferred. Nelson and Winter (1982) point out this variance in procedural skills. Explicit knowledge tends to be adequate when “the pace of the required performance is slow and pace variations are tolerable, where a standardized, controlled context for the performance is somehow assured, and where the performance as a whole is truly reducible to a set of simple parts that relate to one another only in very simple ways.” (p. 82) To the extent that these conditions do not hold, the transfer of the tacit knowledge is likely to be quite important. In such cases, verbal instruction by itself provides only a starting point at best for the acquisition of the skill (Miller, Galanter and Pribram, 1960). For example, the description of how to land an airplane is quite short and could be memorized in a few minutes; however, it is doubtful that the person who memorized this information could actually land a plane, even under ideal weather conditions.

Different kinds of data are more or less vulnerable to tacit knowledge problem. If data were vulnerable to the tacit knowledge problem, documentation would be perceived as insufficient for secondary use, because tacit knowledge cannot be transferred through documentation. **Hypothesis 4: Data less vulnerable to the tacit knowledge problem are perceived as better documented than data more vulnerable to the tacit knowledge problem.**

Qualitative data are more vulnerable to the tacit knowledge problem than quantitative data. There is a well-established tradition of the secondary analysis of quantitative social science data. However, there is a strong skepticism about the secondary analysis of qualitative data. According to Van Den Berg, (2005), knowledge about qualitative data is highly contextual and experience dependent. Data producers gain the knowledge through their direct experience in producing the data. Such knowledge is considered too rich to be conveyed with sufficient detail. Furthermore, much of it relies on the data producers' senses or feelings, which are difficult to articulate. Some researchers are concerned that qualitative data cannot be used sensibly without the accumulated background knowledge that the original investigator acquired during its collection (Blommaert, 1997; Dale, Arber, & Procter, 1988 p. 32). In fact, Van den Berg (2005) argues that complete contextualization of qualitative social science data is unattainable. **Hypothesis 4A: Quantitative data are perceived as better documented than qualitative data.**

When there is a consistent and standard methodology to collect certain kind of data, a lot of knowledge about how the data are collected is common knowledge to many people, and does not need to be transferred from data producers. Therefore that kind of data is less vulnerable to the tacit knowledge problem. Zimmerman (2003) suggested two factors that make data collection simple and easy: the existence and stability of standard methods and low variability in nature. She used survey data as an example of simple data, because surveys have a fairly standard methodology. Boruch (1991, p80) mentioned that survey data are simpler than field experiment data. Birnholtz and Bietz (2000) talked about the difficulties in documenting experimental data. They said that metadata are an inherently incomplete abstraction of what actually took place in an experiment. In HIV/AIDS research, data often cannot be understood without a thorough understanding of the clinical and laboratory processes that produced them. Scientists are not necessarily able to explicate all of the information that is required to understand someone else's work. **Hypothesis 4B: Survey data are perceived as better documented than data collected using other methods.**

McCall and Applebaum (1991) discussed the limitations of codebooks associated with older data. In the past, research methodology was more primitive, some researchers were less meticulous in keeping records, and they did not consider the possibility that their data may be analyzed by later generations (Hyman, 1987). In addition, the need for documented context increases over time as instruments, practices, and standards evolve (Borgman, 2007, p. 230). **Hypothesis 4C: Newer data are perceived as better documented than older data.**

Knowledge Receivers

Incentive

In secondary data use, secondary data users are the knowledge receivers. Users may not want to use secondary data. In some scientific communities, use of secondary data may be looked down upon. Collecting one's own data is often viewed as better than using secondary data. Using secondary data can affect one's chances of getting papers accepted in journals and conferences and can impact the level of respect that is felt toward the author by the community (Birnhotz & Bietz, 2000). This kind of incentive issue affects users' decisions to use secondary data, but is not relevant to perceived documentation quality. Zimmerman (2003) reported uncertainty as another potential cause of unwillingness to use secondary data: "When using data collected by others, it is not always possible to see what one would like to see, a situation that leads to uncertainty."

(p. 160). Different people have different levels of tolerance for uncertainty. Some people would never use secondary data because they can't accept any degree of uncertainty. Some people are willing to accept some types of uncertainty, but not others. Perceived documentation quality is likely to affect the level of uncertainty in using secondary data and therefore affect users' incentives to use secondary data.

Absorptive capacity

The process of knowledge transfer does not end until the receiver assimilates the knowledge. The more capable the receiver is of absorbing the knowledge, the easier the knowledge transfer will be. Such capacity is largely a function of the receiver's preexisting stock of knowledge (Dierickx & Cool, 1989). Existing research uses knowledge distance as an indicator of absorptive capacity (Ivari & Linger, 1999; Tuomi, 1999). Knowledge distance is the overlap or shared knowledge space between the sender and receiver. The more overlap between the sender's and the receiver's knowledge, the shorter the knowledge distance between them, hence the more tacit knowledge the receiver has, and the less explicit knowledge the user needs in order to use the knowledge of the sender (Markus, 2001). Zimmerman (2003) found that the secondary use of ecology data was affected by users' formal knowledge and informal knowledge. Formal knowledge is formal disciplinary knowledge and standards of scientific practice. Informal knowledge is knowledge gained through fieldwork. Borgman (2007, p. 230) further pointed out the effects of knowledge distance between data producers and users: the greater the distance from the data author, the more scientists must rely on documentary evidence. Scholars in an area farther away from the field in which the experiment was conducted, students new to the field, and teachers and policy makers will be constrained in their ability to interpret the available documentation.

Hypothesis 5: The stronger the user's absorptive capacity, the higher the perceived documentation quality.

Knowledge Transfer Channels

I categorize knowledge transfer channels into three types based on the kinds of knowledge that can be transferred through them. One channel for knowledge transfer is the use of documents. Only explicit knowledge can be transferred in this manner. A second channel is interactive conversations, such as face-to-face or phone conversations, meetings, or email messages. Through this channel, a receiver might be able to capture some tacit knowledge through the facial expressions or tones of the sender, but knowledge transferred through this channel is primarily explicit knowledge that is verbalized and not formally documented. When documents are not sufficient to transfer knowledge, conversations may help the receiver to obtain more information or further clarification. A third channel is situated learning. A typical example is an apprentice working with his/her mentor in order to learn craftsmanship not only through language, but more importantly, by observation, imitation, and practice. Tacit knowledge that is very hard to articulate can be transferred through this channel. Sometimes the sender and receiver are not even fully aware that certain knowledge has been transferred. As Collins (2001) describes, "A performs aspects of an experiment in a certain way without realizing their importance; B will pick up the same habit during a visit while neither party realizes that anything important has been passed on."

The depth of involvement of the sender increases from the first to the third channel. The availability and richness of transfer channels affect the difficulty of knowledge transfer. For example, in cases where tacit knowledge plays a key role, but personal interaction between the sender and receiver is not possible, the knowledge transfer is likely to be very difficult. Documentation has never been sufficiently exact to enable replication of the original production of the atomic bomb. However, several other countries created atomic bombs after the United States had done so, based solely on the insufficient documentation, having no direct contact with the people who were involved in the creation of the first atomic bomb. MacKenzie and Spinardi (1995) argue that, in this case, tacit knowledge was re-invented rather than transferred. This re-inventing process was found to be very difficult.

About the knowledge transfer channels for secondary data use, I assume the following: when documentation is adequate, documentation is the only necessary knowledge transfer channel. Otherwise, other channels or sources need to be consulted. Those include (1) other documents, such as previous publications based on the data or websites of data producers; (2) conversations, such as emailing or calling data producers, other secondary users or data archivists; (3) situated learning, such as visiting or even working together with data producers.

Research Questions

The goals of this study are to identify impacting factors of perceived documentation quality, find out how perceived documentation quality affects secondary data use, including users' incentives to use secondary data, and how users overcome inadequate documentation. To achieve this goal, hypotheses formulated based on existing literature need to be tested. In addition, this study tries to identify other impacting factors of perceived documentation quality and characterize the information seeking of secondary data users to overcome inadequate documentation. Answering these questions would inform how to improve documentation quality and help secondary data users. The research questions of this study are:

1. What are the impacting factors of perceived documentation quality? Answering this question includes testing hypotheses proposed before.
2. How does perceived documentation quality affect users' incentives to use secondary data?
3. How do secondary data users overcome inadequate documentation?

Summary

To help identify the impacting factors of perceived documentation quality, documentation of social science data was investigated as a knowledge transfer channel between data producers and secondary data users. A general knowledge transfer model was proposed based on literature in knowledge management, and then applied to characterize the knowledge senders, receivers, knowledge transferred and knowledge transfer channels in secondary data use. In secondary data use, data producers transfer data and knowledge about that data to secondary data users either directly or through an intermediary. Documentation provides explicit knowledge about data. Knowledge receivers are secondary data users. I assume that when documentation is adequate, data producers are the only necessary knowledge senders and documentation is the only necessary transfer channel. When documentation is inadequate, intermediaries and other secondary users may be involved and become knowledge senders as well. Knowledge transfer channels may be expanded to include not only documentation, but also other documented resources and conversations with, and/or situated learning from, data producers, data archivists, and/or other data users. Four possible impacting factors of perceived documentation quality were identified: data producers' incentives and ability, data users' absorptive capacity, the existence of intermediaries between data producers and users, and data's vulnerability to the tacit knowledge problem. Hypotheses about how each factor affects perceived documentation quality were formulated. Research questions of this study were also formulated based on the goal of this study. Besides testing the effects of those impacting factors identified based on existing literature, this study will also try to identify more impacting factors and find out how perceived documentation quality affect secondary data use, including its effect on users' incentive to use secondary data and how secondary data users overcome inadequate documentation.

CHAPTER THREE

METHODOLOGY

The research questions of this study posed several requirements for research methodology. First, the research questions are about *user perceived documentation quality*. Therefore I need to study secondary data users in order for them to report their perception. Second, testing the hypotheses discussed in Chapter two requires quantitative data on a relatively large sample. Third, in addition to hypothesis tests, this study also tries to identify extra impacting factors of perceived documentation quality, and to find out how perceived documentation quality affects secondary data use. This requires qualitative data to explore the details in secondary data use. After evaluating several possible social science research methods, a survey was chosen to collect quantitative data for hypothesis testing, and interviews were chosen to identify extra impacting factors and to study how perceived documentation quality affects secondary data use. The requirement of large sample for hypothesis testing disqualifies several social science research methods that are more suitable for smaller samples, such as in-depth interviews, observation and focus groups. An experiment is another option to collect data on a relatively large sample, but I rejected it two reasons. First, it is very difficult to ask serious data users, such as professors, to commit time for an experiment. Second, in an experiment, users' perception is based on their shallow efforts in a limited time period and they are not motivated to explore the data and documentation as thoroughly as what they do for their own research. Focus groups can be used to meet the third requirement for research methodology. However, focus groups place high requirements on secondary data users. They need to travel to the same location. Even though I may have been able to run the focus group through Internet meetings, participants need to be available at the same time. Another choice is to observe the use process of secondary data users. The strength of this method is: I could have an extended and close contact with secondary data users. However, the length of using a secondary data varies in different situations. Some users use a dataset in one to two year time periods. It is hard to observe a user for such a long time. This method is also intrusive and causes inconvenience to data users: they would have to use the data while I was observing them.

Just as other research methods, surveys and interviews have their limitations. I could not observe more details than what users wrote in the survey or said in the interviews. People sometimes do not remember things clearly or correctly. It is possible that what they said does not match what they really do. Especially in the survey, I could not talk back and forth to explain things if they misunderstood or did not understand the questions. Therefore, questionnaire design is very important. I had to make sure the survey questions are easy to understand, are effective to get information I need. To address these concerns, I did explorative interviews before designing the survey. These interviews helped survey design and helped to answer some research questions.

In the following sections, I discuss how the interviews were conducted and how they helped create and refine the survey instrument, determine the sampling strategy, and define the unit of analysis of the study. Then I will introduce the details of the survey, including survey instrument development, the target population and sampling strategy, pilot survey and formal survey.

Exploratory Interviews

I have experience analyzing the process records[1] and access records[2] of a data archive (Niu & Hedstrom, 2007b). I also used survey data from an archive to practice statistical skills. But those experiences are limited. Lacking effective personal experience of secondary data analysis, to prepare for the formal study, I interviewed 13 secondary data

[1] Process records show how data archivists process a dataset after it arrives the data archive.

[2] Access records show how each data set is browsed online or downloaded.

users. Data users for the interviews were found through the following ways: (1) In 2006, in collaboration with ICPSR and my advisor, I did a survey for the NSF funded project “Incentives for data producers to Provide Archive-Ready Datasets”. Surveyed subjects were all grantees of NIJ who are required to deposit their data into the National Archives of Crime and Justice Data (NACJD) after they finish their projects. Many of them are secondary data users, some of whom left contact information and gave me permission to talk further about data sharing issues. (2) An email asking for volunteers was sent to the mailing list of the International Association of Social Science Information Services and Technology (IASSIST). Several secondary data users on that list volunteered. (3) I contacted some data librarians of several universities. Some of them forwarded my recruiting email to their users. (4) Some users pointed me to other users or group of users. Each of the interviews lasted 30-70 minutes. These interviews were unstructured, qualitative and exploratory.

During the interviews, I asked users to walk me through a secondary data analysis process, talk about the difficulties they encountered in using the data and how they overcame the difficulties, describe their perceptions of the quality of documentation, and discuss how documentation affects their secondary data use, etc. It turned out that some of my questions were not framed in the best way; some of my assumptions were not correct. Findings from the interviews shaped my thoughts and the way I understood and organized existing literature, and helped me to refine the survey questions, answer some research questions and even to identify indicators of certain variables. For example, before conducting this study, I assumed that the role of documentation quality is so important that users would give up using a dataset if it is badly documented. However, it was quite consistent among the interviewees that that is not always the case. So I decided that it was not necessary to study that question again in the survey. I also found that many master’s degree students are not very serious data users for research. Rather they mostly use secondary data for data analysis courses, and therefore the quality of data and documentation are not an important concern for them. I decided to exclude master’s degree students from the sample for that reason. Interview findings also helped the selection of the unit of analysis of the survey. I found many users use one dataset multiple times and many use multiple datasets for one project. If I asked survey respondents to answer a survey based on one research project, they would have to answer questions for each different dataset. This was considered too much a response burden for data users. If I asked respondents to answer the survey based one dataset, they might have used the dataset many times, each time for a different purpose. In that case, their absorptive capacity would change during the process, as they became more familiar with the dataset each time they used it. Therefore I designed a survey in which the unit of analysis was a single use of a single dataset. To make sure respondents had as good recollection as possible, I asked them about their most recent use of a dataset for research. Findings from these interviews will be analyzed together with findings from the future survey.

Survey Instrument Development

To answer the research questions, questions about perceived documentation quality, characteristics of data, users' absorptive capacity, sources and channels where secondary users seek information need to be asked. Below are details about how those questions were designed.

Perceived Documentation Quality

Perceived documentation quality is a core concept in this study. To answer the research questions, there must be some evaluation metrics for evaluating perceived documentation quality. Since no existing metrics were found, a Documentation Evaluation Model (DEM) was created based on existing literature. The DEM evaluates perceived documentation quality on two dimensions: perceived sufficiency and perceived ease-of-use. Sufficiency means the extent to which the documentation provides enough information about the data. It is measured on two aspects: completeness and the overall perception of sufficiency. The indicator "completeness" is borrowed from the Document Quality Indicators (DQI)[3] (Arthur and Stevens, 1989), which is a comprehensive

[3] There are similarities between data documentation and software documentation. Data documentation is important for secondary data analysis, but data producers who produce data for their self-use are often unwilling to document their data for secondary use. They often treat documentation as the last thing to do before the end of their research projects. To help improve

taxonomy for the evaluation of software documentation. In DQI, completeness means all the required components are present in the software documentation. This indicator was changed in two ways. First, instead of using the required components for software documentation, the required components for DEM were decided based on DDI (http://www.ddialliance.org/DDI/dtd/version2-1-all.html?section=2). Second, instead of asking users to check whether each required component "exists or not", I asked users to evaluate how well each required component is described in documentation, which will convey more information than the binary value of "exists or not." The full list of elements defined in DDI is very long. To reduce the burden on survey respondents, I chose elements based on the following rules: (1) choose elements on high levels. DDI has a tree structure. An element may have several levels of sub-elements. For example, about the title of a study, there is a *Title Statement* section, which includes a title element, which again includes a subtitle, an alternative title and a parallel title. The element *Geographic Coverage* has two sub-elements: descriptive text and concept. (2) elements for specific kinds of data were excluded. For example, elements like west, north, east and south bounding longitude, and elements about software used in producing data. A preliminary list was created and tested in the pilot study. After the pilot study, I deleted several elements that two participants had difficulty understanding in the pilot study, such as value labels, variable labels and wild codes. The finalized list includes 19 elements. The values of completeness were the average values of the 19 elements. (See Table 1.)

documentation quality, the social science data archive community proposed the early documentation idea, which basically means to start documentation process from the very early stage of the life cycle of a research project. However, there is evidence showing that that idea is not implemented well by data producers who collect data for their own research. Software documentation has similar attributes. It is very important for software maintenance and use. But software developers often give it a low priority on their deliverable schedule. They tend not to document software in a timely or systematic manner. To help improved software documentation quality, the idea of concurrent documentation was proposed in the software industry, which essentially means: the recording of requirements, design, specification and implementation decisions *as they occur* with the commitment to convey purpose, content and clarity (Tausworthe, 1977). However, "the failure to observe the concurrent documentation principle is common in all application domains" (Arthur and Stevens,1989, p1).

Table 1. The 19 elements for completeness

Title of the dataset	Data collection method	Question text,
Principal investigator(s) of the study	Sample and sampling procedures	Recoded and derived variables
Time period covered by the data,	Weighting information	Frequencies of variables
Geographic location where the data were collected	Response rates	Data file formats
Funding agency/sponsor	Bibliography of publications related to the data	Missing data
Contact person(s) who are responsible for answering questions about the data,	Variables,	-
Purpose and goals of the data collection	Data collection instrument(s),	-

Completeness does capture much of the meaning of sufficiency, but I was not convinced that users necessarily believe that documentation is sufficient simply because each element in the list is well described. Therefore, in addition to completeness, I added indicators for users' general perceptions of documentation sufficiency. General perception of sufficiency means users' overall perception of the extent to which the documentation provides enough information about the data for their purpose of use. This is measured by users' rating of the following three statements: 1) the documentation provided enough information for me to judge the reliability of the data, 2) with the documentation, I did not need additional information about the data for my use, and 3) the documentation provided enough information for my purpose of use. Respondents' evaluations were all based on a 7-point Likert scale.

Another indicator in DQI is usability, which basically means ease-of-use. I used this indicator because existing literature mentioned issues related to the ease-of-use of data documentation (Sieber, 1991). The measures for this construct were created based on the Technology Acceptance Model (TAM) (Davis, 1989). In TAM, measurements for perceived ease-of-use include: (a) easiness to learn, (b) easiness to operate, (c) easiness to get the product to do what the customer wants to do, (d) clarity and understandability of the interaction with the product, (e) flexibility and easiness to become skillful in using the product. I tailored those measures for documentation of data. Since documentation is not operable, measures "b" and "e" were deleted. Measure "c" was made more specific and changed to: easy to find information I need from the documentation. Measure "d" was changed to: the content of the documentation is clear and understandable. An overall measure for ease-of-use was also created. Ease-of-use is measured by how much users agree with the following four statements: 1) overall the documentation is easy to use, 2) it is easy to learn to use the documentation, 3) it is easy to find information in the documentation, and 4) the content of the documentation is clear and understandable. All the four sub-measures were rated by the respondents on a 7-point Likert scale.

Besides ease-of-use and sufficiency, I also created a question about the overall evaluation for perceived documentation quality. The question asks respondents to rate how much they agree with the following statement on a 7-point Likert scale: "The data was well documented."

Sources And Channels Where Secondary Users Seek Outside Information

I designed multiple-choice questions with the "other " option that allows users to provide other sources and channels beyond what were listed in the options provided. The listed options were created based on existing literature. Here are the sources listed in those multiple-choice questions: data producers (people who collected the data), other secondary users of the same data, data archivists, publications based on the data, websites of data producers, websites of data archives and workshops. An "other, please specify" option was provided to allow users to add more sources where they obtained outside information about data. For users who obtained outside information from people, such as data producers, data archivists and other secondary users, channels through which information was obtained were asked. Those channels are: email or telephone, face-to-face conversations and working together. An "other, please specify" option was provided to allow users to add more channels. To investigate the relationship between knowledge transfer channels and the types of knowledge transferred through those channels, and to find out the extent to which tacit knowledge affects users' information seeking behavior, I also asked why users choose certain knowledge transfer channels. The list of reasons provided include: the documentation does not contain that information; the documentation is hard to use; other information sources or channels are immediately accessible; I happen to encounter that information about the data; that information is tacit knowledge that is hard to be documented. An "other, please specify" option was provided to allow users to add more reasons.

User Characteristics

Measurement items for users' absorptive capacity were created based on literature review and preliminary analysis of interview data. Based on findings from Zimmerman (2003) and Borgman (2007, p230), users' professional status and familiarity with the topics of data were created as two initial measures for absorptive capacity. Based on findings from the preliminary analysis of interview data collected for this study, another four items were added to measure users' absorptive capacity: 1) users' experience using exactly the same data, 2) experience using secondary data in general, 3) experience collecting data and 4) experience analyzing self-collected data. Therefore in total, absorptive capacity was measured using six items. Professional status is the professional rank of users, such as professor, associate or assistant professor, post-doctoral researcher, or Ph. D. student. The other five items were measured based on users' ratings of the following statements on a 7-point Likert scale: the topic(s) of the data were within my specialty; I have experience analyzing exactly the same dataset; I have experience analyzing data produced by other people; I have experience analyzing self-collected data; and I have experience collecting research data. Below are the interview findings that provided the basis for constructing the four additional measurements of absorptive capacity.

Experience using the same data makes data and documentation easier to use, and makes users less reliant on documentation. "For me, it is easier to use datasets that are familiar.

If I use serial data that comes out every year, that is easier for me, because I see it on a regular basis (Interviewee #6).” “It (the documentation) is kind of hard to look through. (But) since I am already very familiar with the data, it is easier for me to look through it (Interviewee#7).” “The most recent time I used the data, I didn’t use documentation, because I have used it so many times and I know many things in my head (Interviewee #8). ”

Secondary data analysis experience builds data analysis skills. “I work with secondary data over and over again. I learn through the whole process how to be more efficiently collect the data in the first place, data manipulation, the creation of variables, and selection of variable. Through the process, I become more effective in interpreting results, identifying positive or significant results (Interviewee #5).”

More background knowledge has negative impact on perceived trustworthiness of data. “The less you know, the more you are willing to step on faith. So I think sometimes if somebody made a mistake with their data, if you have been around for a while, you have some experience, you will see those mistakes. Whereas, if it is your first research project. If they tell you that the average sentence was 50 years, are you going to believe it? But if you have been in the business for a while, you know that very few people serve 50 years (in prison). I think in some way more experience make you use the data more easily, but less experience makes you trust the data more (Interviewee #3).”

Data collection experience has a negative impact on perceived data and documentation quality. “I never doubted the quality of data before I did the fieldwork of collecting data for China Statistics Bureau. …. I interviewed an illiterate poor farmer at northwest China. The questionnaire was long. He cared about the incentive money and would like to answer my questions. At the beginning, he thought carefully before answering my questions, but then later, he became tired and impatient, and just answered without thinking (Interviewee #9).”

Borgman (2007), Markus (2001), Ivari & Linger (1999) and Tuomi (1999) all used the knowledge distance between knowledge senders and receivers as an indicator of absorptive capacity. I agree that knowledge distance would be a good measure when the knowledge domain of data producers is easy to define, such as data producers who are individuals. However, when specific data producers are unknown or a big group comprised of people from various disciplines, it is hard to define what is the knowledge domain of the data producers and therefore hard to measure the knowledge distance between the producer and user. The exploratory interviews found that it is often the case that data producers are research groups rather than individuals. In addition, this study only surveys data users. Data producers are not included. I finally decided to measure the knowledge distance between users and data instead of the knowledge distance between users and producers. Users' familiarity with the topics of the data is a measure of the knowledge distance between users and data.

Data Characteristics

The characteristics of data include data collection methods (survey, experiment, etc.), type (qualitative, quantitative, longitudinal, cross-sectional), producer, distributor, and the time of data production(age). Data collection methods, types and data producers were multiple-choice questions. Distributors and age of data were open-ended questions.

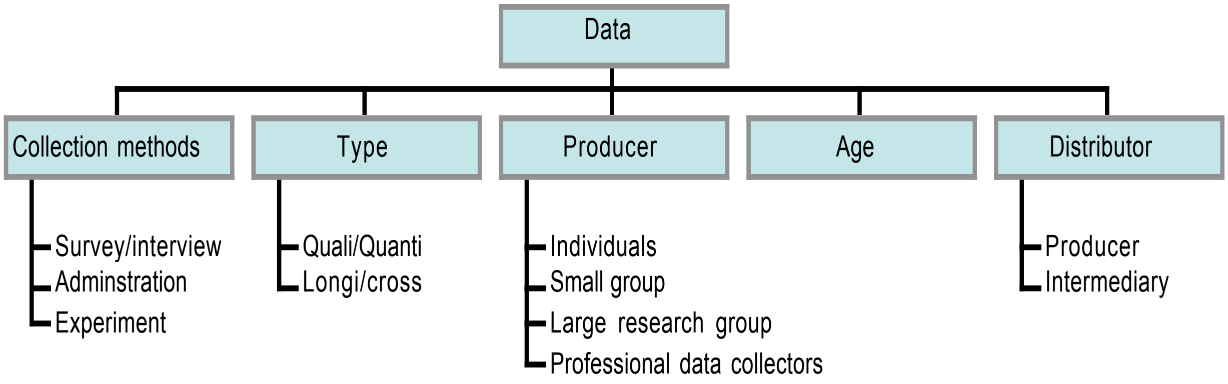

Figure 3. Characteristics of data

I assumed that survey and interview data are two different kinds of data. The former is quantitative, and the latter is qualitative. From the exploratory interviews, nevertheless, I found that many users do not differentiate survey and interview data, because the data they use were survey (quantitative) data collected using face-to-face or telephone interviews. They call those kinds of data survey or interview survey. To avoid confusion, I combined survey and interview as one data collection method.

Target Population and Sampling Strategies

The population of the study is secondary users who use social science data for research. The minimum sample size was determined as 164 users after consulting with professional statisticians. Secondary users for the survey were found through the following ways:

(1) I searched the databases of journal articles of University of Michigan library using the term "secondary data analysis" in the "Social Science" category and "Business/Economics" category. I selected publications based on secondary data analysis after 2002, and then searched for the authors' email addresses. I collected about 150 users this way.

(2) I collected secondary data users from the websites of data archives. ICPSR is the world's largest archive of digital social science data. It provides information about publications based on datasets they archive. I collected about 200 authors who have publications based on data archived at ICPSR since 2002. I also collected a list of 479 ICPSR Official Representatives (OR) from the website of ICPSR. ORs are representatives of ICPSR member institutions. There are two types of ORs. One group of them are professors or researchers who are secondary data users. Another group of ORs are data librarians. They help secondary data users locate and use data, but they do not use secondary data for research. Qualitative data are used rarely compared with quantitative data. To make sure my survey responses include qualitative data users, I searched the websites of the world's biggest qualitative data archive: the Economic and Social Data Service (ESDS) Qualidata. For each archived dataset, ESDS Qualidata provides a list of publications based on that data, separated into two parts: publications by

the Principal Investigator of the data and publications based on secondary data analysis. I collected the names of those secondary users, and searched online for their email information. Besides the publications, the website of the Qualidata also provides information about who are current users of their data, even though they have not published any work based on the data yet. The website also points to some other articles about secondary analysis of qualitative data. I found the authors of those articles who are secondary users of qualitative data.

(3) I also asked secondary data users to forward my recruitment email to other users.

Pilot Survey

Part of the survey instrument was pre-tested in a related study. In collaboration with the National Archive of Crime and Justice Data (NACJD) at ICPSR, my advisor and I designed a usability study for data and documentation. In the study, a group of subjects were asked to do some exercises and finish a questionnaire based on the data and documentation in the condition it was in before being processed by the data archive. Another group of subjects were asked to do exactly the same thing based on the same data and documentation after it had been processed. The purpose of the study was to find out the differences between using pre-processed and processed data. We expected to find that subjects using processed data could finish the exercises faster, would find the exercises easier to do, obtain more precise results, and rate the documentation as easier to use and more sufficient. I tailored the survey instrument for my dissertation proposal to fit that study. Data archivists at ICPSR helped me refine the question texts. The tailored survey instrument was pre-tested in the pilot study conducted in March 2008. Those questions worked well in the pilot study. None of the 4 pre-testers complained about those questions.

In May 2008, I hired two graduate students at the University of Michigan to test my survey instrument. One is a Ph. D student at the School of Public health. He has much experience using large quantitative datasets. Another one is a master student from the School of information. She filled out the survey based on a small qualitative dataset that she used while she worked in a company. Both of them were native English speakers. That pilot study was conducted in person. I observed them while they filled out the online survey. After they finished the survey, I asked them questions related to clarity of the questions, the ordering and sequence of questions, grammar errors, etc. I changed the wording and order of the questions. I also deleted some items for the completeness of documentation that were not easy to understand.

Formal Survey

The formal survey started at the end of May 2008. On May 29, I sent out the first wave of surveys to the 758 secondary data users. Two weeks later, I sent out the second wave of surveys to those who did not reply. Another two weeks later, I sent out the survey to the list of ICPSR Official Representatives (ORs). Some of them are secondary data users, some of them are data librarians who tend to have connections with secondary data users. I used two different invitation emails to the two groups of ORs. For the group who are secondary data users, I used the same invitation email as the one I used for other secondary data users. For this group of ORs and other secondary data users, I could track down who replied the survey and who did not, so I sent out another wave to people who did not reply after two weeks. For the group who are data librarians, I used a different invitation email. In the email, I asked their favor to forward my email to secondary data users. To avoid giving them the impression that I only survey ICPSR users, I did not say anything about their relationship with ICPSR, nor did I say anything about ICPSR data. Some people figured that I email them because they are ICPSR users and emailed me back for confirmation, I replied saying that I need secondary data users, not only ICPSR data users. Some people forward my email immediately, some posted my survey link to their web site. No incentives cash were applied during the whole process of this survey. Below is a table showing how the survey was administered.

A total of 1,260 surveys were sent out. By August 31, 2008, 431 surveys were received in total. Some respondents are non-social science data users, for example, dentistry data users, biology data users. Some data are used for non-research purposes, for example, for librarians' reference services, decision making, reporting, program evaluation, etc. Some were used a very long time ago, such as in 1990 and 1998. After removing those responses, there are 384 usable responses. The survey collected both highly structured quantitative data and open-ended qualitative data. Each respondent was asked about his or her most recent experience using a dataset produced by other people.

Summary

A survey and interviews were chosen as the data collection methods. Survey data will be used for hypothesis testing. Interview data will be use to identify extra impacting factors of perceived documentation quality and find out how perceived documentation quality affects secondary data use. Interviews were conducted before the survey. Preliminary analysis of the survey helped the survey design, such as deciding the survey population and unit of analysis, identification of indicators of perceived documentation quality, etc. To evaluate perceived documentation quality, a Documentation Evaluation Model was constructed based on existing literature. In the model, perceived documentation quality was evaluated on two dimensions: sufficiency and ease-of-use. This model was used as the basis for creating survey questions about perceived documentation quality. The survey population was defined as secondary users who use social science data for research. The unit of analysis was a single use of a single dataset. Each secondary data user was asked about his/her most recent use of a particular dataset. The survey was pre-tested in a related study, and then pre-tested again with two data users in person. A total of 1,260 surveys were sent out. 431 surveys were received in total, and 384 surveys were usable responses.

CHAPTER FOUR

FINDINGS AND DISCUSSIONS

In this chapter, I will first discuss the survey responses collected for this study. Then I will answer research questions based on the combined analysis of interview and survey data. There are five parts in this chapter: (1) sample characteristics; (2) the reliability and validity of DEM; (3) impacting factors of perceived documentation quality; (4) the effect of perceived documentation quality on users' incentive to use secondary data; (5) the information seeking of secondary data users.

Sample Characteristics

The population for this study is people who use secondary data to conduct social science research. Unlike other populations, such as female faculty members in the United States, this population cannot be defined based on simple demographics. Unlike some other populations, such as all users of Facebook, there is not a single registry for all secondary data users for social science research. It is hard to define exactly what the population of secondary data users looks like, except for several characteristics. First, secondary data users should come from various social science fields. Second, their age, education and occupation should allow them to conduct serious research. The sample of survey responses indeed satisfies these two criteria. Survey respondents are from various social science fields. They are from the sub-fields or interdisciplinary studies of sociology, economics, psychology, public health, political science, public administration, education, business, geography and demography. Those sub-areas are aging, gerontology, epidemics, criminology, family studies, social capital, gender equality, social capital, social work, housing, finance, migration, communication and religion. Users are between 22 and 74 years old. The mean age is 42 years old, the standard deviation is 11.6 years. Their highest degrees are mostly doctoral degrees (72%) including Juris Doctor, Doctor of Medicine and Doctor of Philosophy degrees. There are a few other degrees such as Diploma from countries outside of the United States, and some doctoral candidates without the dissertation. Respondents are predominantly (93%) from universities and colleges. The remainders are from various research centers, government agencies, etc. The table below shows the professional status of survey respondents. In raw data, the "others" category includes researchers, research fellows, research professors, professor emeritus, research economists, data managers, IT staff, research scientists, data analysts, undergraduate students, etc. The professional ranks in the "other" category were separated into other categories. For example, "professor emeritus" was re-assigned into the full professor category. Professional status is a categorical variable, but was recoded into an ordinal variable. (See Table 2 below.)

Table 2. Users' professional status

Raw			Recoded		
Professional Rank	Numbers	Percentages	New codes	Numbers	Percentages
Full professors	79	22.1%	7	83	23.2%
Associate professors	68	19.0%	6	70	19.6%
Assistant professors	65	18.2%	5	74	20.7%
Post-docs	20	5.6%	4	36	10.1%
Doctoral candidate	50	14.0%	3	50	14.0%
Doctoral pre-candidate	14	3.9%	2	14	3.9%
Masters	28	7.8%	1	28	7.8%
Others	31	8.7%	-	-	-
-	-	-	0	3	0.8%
Total	358	100%		358	100%

This is not a random sample and I did not intend to make it a random sample. Since qualitative data are less commonly used than quantitative data, a random sample would have produced too few qualitative data users to test hypothesis 4A. I did purposive sampling to make sure I included enough qualitative data users. As expected, I did get a small number of survey responses based on the use of qualitative data, but enough to test hypothesis 4A. 29 (15%) [1] users used qualitative data. 167 (85%) used quantitative data. 76% of respondents used survey/interview data, 5.5% used census data, and 11.6% used administrative records. Very few survey respondents used experimental data, data collected by observing the behavior of human subjects, or historical records. These low percentages reflected the reality that those kinds of data are not commonly used as secondary data for social science research. Since I did not have specific hypotheses about experimental, observational and historical data, I did not use purposive sampling to make sure I included enough users of those kinds of data.

Secondary data analysis was defined as "the analysis of data for a different purpose than what the data were originally collected for, possibly by the original data producers themselves, or in collaboration with other people, or by entirely different people." In this study, I intended to study only the cases in which people use data produced by other people, with or without collaboration with data producers. The sample met my expectations. All the respondents used data produced by other people. 294 (77%) of them reported they did not collaborate with data producers in using secondary data. 89 (23%) of them reported that they did.

Enough responses were collected to test hypothesis 3. Among users who provided precise information about where they obtained data[2], 74% are from data producers, 26% are from intermediaries such as data archives, including ICPSR, United Kingdom data archives and various other data centers. Many other users obtained data from colleagues and advisors. Some users gave only a person's name for the source of data. It is hard to tell whether those colleagues, advisors or other people are data producers, secondary data users or intermediaries. Some obtained data from their local institutions, such as their lab, where they are employed, libraries, commercial vendors, etc. 17% of users obtained data from ICPSR, which was the most common intermediary. Most (96%) data were produced

[1] Even though the number of total usable responses is 384, only 196 users answered whether the data were qualitative or quantitative data. This happens to other questions as well. Therefore the sample size (N) changes for different variables.

[2] 84(22%) users did not give exact information about where they obtained data. They said they get data from online, Internet, www etc.

after 1990.

Enough responses were collected for testing hypotheses 1 and 2. Table 3 below shows the distribution of different data producers. Respondents in the "don't know" category were treated as missing data (no answer) in the analysis. Producers in the "other" category were recoded and separated into other categories.

Table 3. Types of data producers

Data producers	Numbers	Percentages
Individual researchers	13	3.5%
Small research groups (2-5 people)	29	7.8%
Large research groups (6 or more people	94	25.3%
Research organizations	39	10.5%
Government organizations	155	41.7%
Don't know	14	3.8%
Others	28	7.5%
Total	372	100%

240 (67%) of users are affiliated with organizations in the US. Other users are from 14 European countries (in the order of number of respondents, from the most to the least, Switzerland, UK, Italy, the Netherlands, Germany, France, Spain, Austria, Belgium, Denmark, Russia, Slovenia, Sweden), two Asian countries (Japan and Singapore) and two South American countries (Mexico and Brazil), Australia, and Israel. Consistent with the above, datasets are also produced in different countries including by international organizations. Different countries have different data sharing policies, which may affect data producers' incentive to documentation data. It is possible that data produced in countries with more strict data sharing policies are better documented than data produced in countries with loose data sharing policies. Therefore, when comparing data produced for sharing with data produced for self-use, I not only ran statistical analysis on all datasets, but also ran the same statistical tests on datasets produced only in the United States only. I will compare results from the two sets of tests and see if there are any differences.

Users were asked about their most recent use case of a dataset. Just as expected, in the survey responses, many datasets were used by very few users and few datasets were used by many users. See Figure 4.

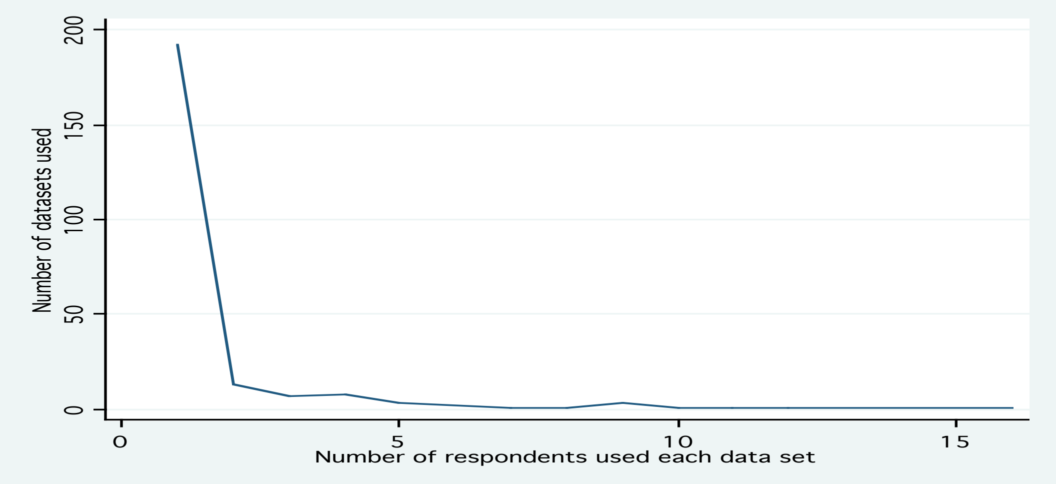

Figure 4: Skewed distribution of the use of datasets

Table 4: Most frequently used data sets

Title of dataset	Number of respondents who used the dataset
Census	16
National Longitudinal Study of Adolescent Health	12
General Social Survey	11
National Longitudinal Survey of Youth	10
European Social Survey	9
National survey of families & households	9
Swiss Household Panel	9
World value survey	8
National Health Interview Survey	7
Current Population Survey	5
German Socio Economic Household Panel	5
Uniform Crime Reports	5
British Household Panel Study	4
Correlates of War	4
Eurobarometer	4
Health and Retirement Study	4
National Education Longitudinal Study	4
National Election Studies series	4
National Health and Nutrition Examination Survey	4
World Development Indicators	4

The Reliability And Validity Of DEM

Perceived documentation quality is a core construct in this study. A Documentation Evaluation Model was proposed as a way to evaluate perceived documentation quality. Since the model was newly created for this study, before using it to conduct analysis, it was important to check the reliability and validity of this model. In this section, I will discuss both the quantitative and qualitative data that I used to check the reliability and validity of this model.

Ease-of-use

As discussed earlier, four items were created to measure the ease-of-use of documentation. Each item was measured based on users' rating of how much they agree with a statement. The ratings were on a 7-point Likert scale. Each of the four items has a highly skewed distribution. This means statistical methods that require normal distribution cannot be used in data analysis.

Table 5. Value distributions of the four items for ease-of-use

	It was easy to learn to use the documentation		It was easy to find information from the documentation		Documentation content was clear and understandable		Overall speaking, the documentation was easy to use	
Ratings	Freq.	Percent.	Freq.	Percent.	Freq.	Percent.	Freq.	Percent.
1	5	1.4%	9	2.5%	7	2.0%	9	2.5%
2	18	5.1%	25	7.0%	14	3.94%	16	4.5%
3	22	6.23%	28	7.9%	23	6.5%	20	5.7%
4	30	8.5%	39	11.0%	37	10.4%	31	8.8%
5	57	16.2%	61	17.1%	57	16.1%	58	16.4%
6	91	25.8%	88	24.7%	101	28.5%	97	27.4%
7	120	34.0%	102	28.7%	109	30.7%	116	32.8%
N/A	10	2.8%	4	1.1%	7	2.0%	7	2.0%
Mean	5.53		5.24		5.48		5.50	
Standard Deviation	1.56		1.68		1.54		1.59	

There are strong correlations between the four items of ease-of-use, which is a sign of strong convergent validity according to the Multi-Trait Multi-Method (MTMM) method. All the correlations are significant (p=0.00).

Table 6. Correlations between the four sub-measures of ease-of-use

	Easy to learn to use	Easy to find	Content clear and understandable
Easy to find	0.81*	1	-
Content Clear and understandable	0.81*	0.81*	1
Overall easy	0.82*	0.83*	0.87*

*: $p<0.01$

The Cronbach's alpha of ease-of-use is 0.95, which is much higher than the minimum threshold of .7 suggested by Nunally (1978) for social science research. Following the suggestion of Carmines and Zeller (1979), a factor analysis was then performed on these four measures and they were found to load onto one factor (N=340[3], eigenvalue: 3.47, explained 87% of the variance in the data). The factor loadings of the four items were all above 0.90 (Easy to learn to use the documentation: 0.92; Easy to find information from the documentation: 0.93; Content of the documentation is clear and understandable: 0.94; Overall, the documentation was easy to use: 0.94)

Sufficiency

The sufficiency of documentation was measured by four variables as well. Completeness is the average value of its 19 measurement items from the DDI, each of which is the rating of how well an element is described in documentation. (See Tables 28 and 29 in Appendix A for the value distributions of the 19 elements.) Each of the other three variables was the respondents' rating of how much they agree with a statement. Those ratings were on a 7-point Likert scale. Each of the four measures for sufficiency has a highly skewed distribution. This means statistical methods that require normal distribution cannot be used in data analysis.

[3] Due to missing data, this N is different from the total usable responses.

Table 7. Value distributions of the three items for sufficiency

	With the documentation, I did not need additional information about the data for my purpose of use.		The documentation provided enough information for my purpose of use		The documentation provided enough information for me to judge the reliability of the data.	
Ratings	Freq.	Percent.	Freq.	Percent.	Freq.	Percent.
1	31	8.7%	10	2.8%	10	2.8%
2	44	12.4%	14	4.0%	12	3.3%
3	40	11.2%	17	4.8%	21	5.8%
4	41	11.5%	26	7.3%	43	11.9%
5	47	13.2%	52	14.7%	61	16.9%
6	64	18.0%	79	22.3%	73	20.3%
7	83	23.3%	148	41.8%	127	35.3%
N/A	6	1.7%	8	2.3%	9	2.5%
Mean	4.58		5.67		5.48.	
Standard Deviation	2.02		1.61		1.61	

The four measures of sufficiency are moderately correlated. All correlations are significant (p=0.00).

Table 8. Correlations between the four items of sufficiency

	Completeness	Sufficiency for reliability judgment	No need to seek additional information
Sufficiency for reliability judgment	0.59*	1	-
No need to seek additional information	0.47*	0.56*	1
Documentation provided enough information for use	0.55*	0.58*	0.65*

*: p<0.01

The Cronbach's alpha of sufficiency is 0.83. Factor analysis showed the four measures load onto one factor (N=338, eigenvalue: 2.70, explained 68% of the variance in the data). The factor loadings were around 0.80 (No need to seek additional information, 0.82; documentation provided enough information, 0.85; completeness: 0.79; Sufficiency for reliability judgment: 0.83)

Overall Quality

In addition to ease-of-use and sufficiency, another variable was created for a general evaluation of how well the data were documented. This was measured based on users' rating of how much they agree with the statement "overall speaking, the data is well documented." The higher the value, the better the documentation is perceived. The mean of these values is 5.65. The standard deviation is 1.63.

Data show strong correlations between sufficiency, ease-of-use, and the overall measure of perceived documentation quality.

Table 9. Correlations between the three measures for perceived documentation quality

	Overall quality	Ease-of-use
Ease-of-use	0.75*	1
Sufficiency	0.79*	0.76*

*: p<0.01

The significant and strong correlation between the two measures shows more sufficient documentation tends to be easier to use.

Qualitative data from the survey and interviews shed more light on the validity of the model. Consistent with the DEM, we found that most of the problems users encountered with documentation fall into two categories: sufficiency and ease-of-use. Many users complained about the absence of certain elements or incomplete descriptions, which supports completeness as a valid measure. Some data have very limited documentation, such as when only variable names or a questionnaire are provided. Some documentation is very brief and tersely written. Examples of information missing from the documentation include the meaning, measurement, construction, and source of variables, methods used for recoded and derived variables, the process of collecting and entering data, sample design, response categories, response rates and anomalies in responses, instrument reliability, limitations of data, information about the purpose of the study, links between variables and the data collection instrument, links between different levels of data, the time period covered, details of experiment design, the reasons for missing data, imputation of missing data, the rationale for changing measures between waves, relationships between a dataset and other data collected by the same producer, coding categories, skip patterns, etc.

Respondents used the following terms to describe documentation that is hard-to-use: irritating, annoying, confusing, cumbersome, complicated, fragmented, dispersed, not user-oriented, not transparent, puzzling, long, huge, massive, unorganized. Also consistent with the DEM, problems related to ease-of-use can be categorized into the following categories: (1) Hard to find the information needed because documentation was available only in hard copy, information was dispersed across multiple files, no cross-references were provided between various parts of the documentation, an unorganized and overwhelming amount of information, or the documentation and data are stored in different places. (2) Hard to understand because descriptions were too tersely written, the terminology was not always clear, and scanned codebooks were blurry and very difficult to read. No respondents mentioned difficulties in learning to use documentation. Even though statistically the measure "easy to learn to use documentation" has high correlations with other measures, which is a sign of construct validity, I decided it was not necessary and should be dropped from the model. After removing "easy to learn", the Cronbach's alpha of ease-of-use is still 0.94. The remaining three sub-measures still load onto one factor. In this dissertation, I will only use the remaining three sub-measures for hypotheses testing.

Besides sufficiency and ease-of-use, users mentioned the accuracy of the documentation as another type of problem. Users often detect errors in documentation based on the inconsistencies between data and documentation. For example, questions of accuracy arise when response values do not match the skip patterns in the questionnaire, when weights in the data do not match weights in the documentation, and when the electronic and hard copy versions of the documentation are different. This finding is consistent with the DQI, where consistency was used as a measure for the accuracy of software documentation. This problem was anticipated before the DEM was constructed. Accuracy was not included in the model because of two reasons: first, the consistency problem is very closely related to data. In other words, if a user detects inconsistency between data and documentation, sometimes it is hard to tell whether the data are wrong or the documentation is wrong. Since DEM was constructed only for documentation, accuracy was not included. Second, it is hard for secondary users to detect errors in documentation other than inconsistency. Because secondary users are not involved in collecting data, it is reasonable to assume that they have to believe in whatever was delivered to them unless some inconsistencies occur. If a future study was conducted to evaluate the quality of both data and documentation, accuracy needs to be included for a more complete evaluation.

Impacting Factors Of Perceived Documentation Quality

In this section, I will use quantitative data to test the hypotheses proposed earlier. In addition, based on qualitative data, I will try to identify more impacting factors, or more details about how each factor affects perceived documentation quality. The none-parametric Mann-Whitney U test was used to test hypotheses. T-test was considered but the assumptions of normality and equal variance were violated.

Hypothesis 1: Data produced for sharing are better documented than data produced for self-use.

Data were divided into two categories based on the type of data producer. Data produced for sharing include data produced by survey research organizations, commercial data companies who sell data for profit, and some government agencies that have a tradition of producing and sharing data, such as the Census Bureau, Bureau of Labor Statistics, and the Center for Disease Control. Data produced for self-use include data produced by individual researchers and small research groups for self-research, or by some organizations as by-products of their business process. **Documentation of data produced for sharing is more sufficient (z=-3.3, p=0.00), easier to use (z=-2.7, p=0.01), and of better overall quality (z=-3.8, p=0.00) than data produced for self-use.** The difference remains the same when statistical tests were conducted for data produced only in the United States. This difference is consistent with qualitative interview data. As interviewee #5 said: “Data sets that are easy to use are often produced for the purpose of distribution, the hardest data to use are data that aren't designed for other people to study.

Table 10. Difference in overall quality between data produced for self-use and data produced for sharing

	For self-use		For sharing	
Ratings of Overall Quality	Frequency	Percentage	Frequency	Percentage
1	5	7.0%	2	0.8%
2	3	4.2%	5	2.1%
3	4	5.6%	11	4.6%
4	9	12.7%	16	6.7%
5	16	22.5%	25	10.5%
6	12	16.9%	63	26.5%
7	22	31.0%	116	48.7%
Total	71	100%	238	100%
Average ratings	5.1		6.0	

Ease-of-use

The ease-of-use measure is the average value of its three components. Because the results of averaging the values of the three components were not always integers, to make the tabulation short, I rounded the averages to the closest integers. For example, the number 2.33333 was rounded to 2. The number 2.66667 was rounded to 3.

Table 11. Difference in ease-of-use between data produced for self-use and data produced for sharing

	For self-use		For sharing	
Ratings of Ease-of-Use	Frequency	Percentage	Frequency	Percentage
1	3	4.3%	2	0.8%
2	3	4.3%	9	3.8%
3	7	10.0%	17	7.2%
4	7	10.0%	18	7.6%
5	19	27.1%	43	18.1%
6	15	21.4%	71	30.0%
7	16	22.9%	77	32.5%
Total	70	100%	237	100%
Average ratings	5.1		5.6	

Table 12. Difference in sufficiency between data produced for self-use and data produced for sharing (same rounding procedure for ease-of-use were conducted)

	For self-use		For sharing	
Ratings of Sufficiency	Frequency	Percentage	Frequency	Percentage
1	1	1.4%	2	0.8%
2	6	8.3%	3	1.2%
3	6	8.3%	16	6.6%
4	15	20.8%	27	11.1%
5	17	23.6%	54	22.2%
6	16	22.2%	74	30.5%
7	11	15.3%	67	27.6%
Total	72	100%	243	100%
Average ratings	4.8		5.5	

Interview data revealed more details about how some data produced for self-use were poorly documented. Interviewee #2 complained about problems in using data produced by other individual researchers. She said the only documentation provided for a dataset she used was variable names. She did not know what types of questions were asked in collecting data, how the questions were asked, or how the questionnaires were administered. It was hard to differentiate similar variables. She described the use process as a hassle. She would have preferred to collect her own data, but she used the data because she was assigned to do so. Interviewees also described administrative records directly obtained from government agencies as messy and poorly documented. Interviewee #1 mentioned his difficulty using a set of police records where police officers used the address of the parking lot at the police station for the residential address of homeless prisoners without including this coding practice in the documentation. When a secondary user analyzes the data and tries to identify hot spots for crime, the police station appears to be a hotspot. This user didn't know that is an error until someone from another police department told him. Those errors are very common in police records. If users trust the data without further investigation, wrong conclusions can be drawn.

Consistent with existing literature (Borgman, 2007, p. 167; Markus, 2001), this study found that the unawareness of users' needs challenges data producers' ability to document data well for secondary use, which may be a more serious issue for producers who produce data for self-use. Interviewee #3, who is also a data producer, said:

> “The real difficulty is you are the data collector and you are very familiar with the data, it is hard to realize that somebody else won’t recognize something, or doesn’t know something, you don’t know what you are missing when you document it.”

Also consistent with existing literature (Wicherts, et al., 2006), incentive issues cause poor documentation of data produced for self-use. Interviewee #3 mentioned the incentive issues of data producers who produce data for self-use: “For some people, there is little incentive. It’s kind of one of the last things that you do with the project. You kind of rush into and get it over with.” In addition to reasons mentioned in existing literature, this study revealed another cause of insufficient documentation of data produced for self-use: data producers’ perception of their data sharing responsibility affects the way they document data. Interviewee #3 thought his responsibility was only to share and document raw data, but not to share and documented processed data:

> “In my experience, normally If you analyze data, you do transformations, you might have individual items scores, you run a factor analysis, you come up with scale, so you might take 20 individual items and turn them into 3-4 scales, when I submitted data to the archive, I tend to give them raw data. So I don’t give them scale. They may or may not want that, In some way it is not my job. My job is to give them data so that they can work with it.”

Interviewee #1 talked specifically about the incentive issues of administrative records: many government agencies are not motivated to document data well because they do not need high quality documentation for their self-use. Some government agencies use data for very simple purposes, so they can’t see some errors in the data and documentation. Even if the data analyst at a government agency saw the problems with data, political issues can become a barrier to improve data and documentation quality.

> “Unless the analyst get support from high up from supervisor, there are only so much they can do. Historically, the data …..were not a high priority. In the motor vehicle crash unit in the police department, (the data) is very often in the basement of the building.”

Interviewee #1 also talked about another situation where high quality data and documentation is not what some government agencies want: “Sometimes police departments may reduce the number of reported crimes to reduce their crime rates and hence damage the reliability of data.”

Data sharing requirements and more advanced use of data can reveal problems in data and push for the improvement of data quality. Interviewee #1 said:

> "When you start to plot the data on the map, you start to see problems…. Only in the last 10 years, most of them (police departments) have started to realize that the data got a lot of noise in it." "The U.S. Department of Justice has been pushing to improve data quality for many years. There are several dimensions that they are pushing: 1) trying to get jurisdictions to share their information systems with each other in order to build a metropolitan-wide crime information system; 2) trying to get police departments to use GIS more extensively in their analysis as way of shaping their intervention policies; and 3) trying to get police departments to conduct more sophisticated analyses." "Many of the federal agencies have really raised the priorities to improve the quality of the information…… The status is just slowly improving."

It is very consistent across all interviewees that survey data produced for sharing are well documented. Even though administrative records directly obtained from producers are generally messy and poorly documented, some administrative records are compiled and shared with the public. Those administrative records become well documented as well, This is consistent with what interviewee #1 said: data sharing requirement pushes government agencies to improve data and documentation quality. For example, the "Federal Court Cases: Integrated Data Base, 1970-2000" is compiled from 94 district and 12 appellate court offices throughout the United States (http://webapp.icpsr.umich.edu/cocoon/ICPSR-STUDY/08429.xml). The Uniform Crime Reports (UCR) are produced from data provided by nearly 17,000 law enforcement agencies across the United States (http://www.fbi.gov/ucr/ucr.htm). Users often obtain these compiled data records from some intermediary organizations, such as data archives, data companies, etc. Interviewee # 6 mentioned that administrative records obtained from ICPSR were very well documented, but administrative data obtained from state agencies were usually poorly documented.

Datasets produced by professional data producers for sharing are often large and complex. Large datasets are characterized by larger sample size (or population in terms of census data) and more details about each unit of analysis (such as more variables for quantitative data, more detailed questions for qualitative data). The complexity is about the structure of datasets. There is a hierarchy in some datasets. For example, a group of professors collected a dataset that combines survey data and court records. That dataset covers 390 court cases. It includes 5 subsets. One subset includes court records for each case. Each of the other four subsets includes interviews with jury members, judges, prosecutors and defendants for each case. In some census data and survey data, sometimes there are records for each household, and then records for each member of the household. If the data is also longitudinal, use of the data would be even harder, because users need to match households and family members across waves. Sometimes variables or survey questions change across the waves, users also need to match different variables across waves. However, complex datasets are hard to use but not necessarily poorly documented, because they often are produced by professional data producers.

Hypothesis 2: Data produced by large producers are perceived as better documented than data produced by single researchers or small research groups.

No significant difference was found between data produced for by large research groups and data produced by individuals and small groups.

Hypotheses 3: Data distributed by intermediaries are better documented than data distributed by data producers.

Statistics based on survey data suggested that **intermediaries (including data archives and various data centers) are effective in improving the documentation quality of data produced for sharing.** Mann-Whitney U test on all data points did not find any significant difference between data distributed by intermediaries and data distributed by data producers. However, documentation of data produced for sharing and distributed by data archives are more sufficient (z = -1.8, p=0.07), easier to use (z = -2.0, p=0.05) and has overall better quality (z = -1.8, p=0.08) than data produced for sharing and distributed by data producers. There is no significant difference between data produced for self-use and distributed by intermediaries and data produced for self-use and distributed by data producers. The insignificant result may be caused by the very small sample size of this kind of data. There were only eight responses belonging to the category: produced for self-use and distributed by intermediaries. This is quite reasonable:

if people produced the data for self-use, it is unlikely that they will ask a data archive to distribute their data.

Table 13. Difference in ease-of-use of documentation for data produced for sharing

	Produced for sharing and distributed by producers		Produced for sharing and distributed by intermediaries	
Ratings of Ease-of-use	Frequency	Percentage	Frequency	Percentage
1	1	0.9%	-	-
2	4	3.5%	-	-
3	9	8.0%	4	8.7%
4	11	9.7%	3	6.5%
5	21	18.6%	7	15.2%
6	30	26.6%	11	23.9%
7	37	32.7%	21	45.7%
Total	113	100%	46	100%
Average ratings	5.5		5.9	

Table 14. Difference in sufficiency of documentation for data produced for sharing

	Produced for sharing and distributed by producers		Produced for sharing and distributed by intermediaries	
Ratings of Sufficiency	Frequency	Percentage	Frequency	Percentage
1	2	1.7%	-	-
2	2	1.7%	1	2.2%
3	9	7.6%	-	-
4	16	13.5%	4	8.7%
5	30	25.2%	12	26.1%
6	27	22.7%	12	26.1%
7	33	27.7%	17	37.0%
Total	119	100%	46	100%
Average ratings	5.4		5.8	

Table 15. Difference in overall quality of documentation for data produced for sharing

	Produced for sharing and distributed by producers		Produced for sharing and distributed by intermediaries	
Ratings of Overall Quality	Frequency	Percentage	Frequency	Percentage
1	2	1.8%	-	-
2	4	3.5%	1	2.17%
3	5	4.4%	1	2.17%
4	10	8.8%	1	2.17%
5	14	12.3%	5	10.87
6	28	24.6%	12	26.1%
7	51	44.7%	26	56.5%
Total	114	100%	46	100%
Average ratings	5.8		6.3	

Table 16. Difference in overall quality of documentation for data produced for self-use

	Produced for self-use and distributed by producers		Produced for self-use and distributed by intermediaries	
Ratings of Overall Quality	Frequency	Percentage	Frequency	Percentage
1	3	6.7%	-	-
2	1	2.2%	-	-
3	3	6.7%	-	-
4	7	15.6%	1	12.5%
5	10	22.2%	1	12.5%
6	8	17.8%	2	25%
7	13	28.9%	4	50%
Total	45	100%	8	100%
Average ratings	5.1		6.1	

Qualitative data from the interviews revealed more details about how helpful intermediaries are in improving data and documentation quality. Interviewee #6 said: "ICPSR does fantastic job. Data from ICPSR or other data archives are absolutely easier to use than data from individual researchers." Interviewee #7 calls a data archive for various help periodically. She attended data analysis classes offered by the data archive, asked them to review papers, asked questions about the release of new data, and also the meaning of some variables. She gave one example to show how helpful the data archive is:

> "For one variable that I have been using for years, I misunderstood how it was measured. Until last summer when I talked with the data archive, it became clear to me. I thought I knew the meaning of it. But I was wrong."

Interviewee #3 pointed out the limitation of data archives in helping improve data and documentation quality. He believes that documentation guidelines are helpful for data producers to improve documentation quality, but they can't solve all the problems of using data produced by other people.

> "It would help. It can't hurt. I am not sure in specific. I think in each specific case, it is going to be a problem, because the people who write the guidelines don't know the specific data set. So they write general guidelines, it will improve things. But it's not going to fix the problem in any given data set."

This is consistent what Zimmerman (2003) pointed out: the work of intermediaries has to build on that done by data producers. Data producers should take the main responsibility for preparing data for sharing.

Hypothesis 4: Data less vulnerable to the tacit knowledge problem are perceived as better documented than data more vulnerable to the tacit knowledge problem.

The vulnerability to the tacit knowledge problem was not quantitatively measured. Instead, several categories of data were considered as less or more vulnerable to the tacit knowledge problem based on the description in existing literature (Zimmerman, 2003) (Van Den Berg, 2005).

Hypothesis 4A: Quantitative data are perceived as better documented than qualitative data.

Documentation for quantitative data is perceived as more sufficient ($z = 1.81$, $p = 0.07$) and easier to use ($z=-1.7$, $p=0.10$) than documentation for qualitative[4] data. There was no significant difference in overall quality.

[4] When datasets include both qualitative and quantitative data, they were coded as qualitative data.

Table 17. Difference in sufficiency between qualitative and quantitative data

	Qualitative data		Quantitative data	
Ratings of Sufficiency	Frequency	Percentage	Frequency	Percentage
1	1	3.7%	1	0.6%
2	1	3.7%	7	4.3%
3	3	11.1%	10	6.1%
4	6	22.2%	21	12.9%
5	8	29.6%	39	23.9%
6	3	11.1%	50	30.7%
7	5	18.5%	35	21.5%
Total	27	100%	163	100%
Average ratings	4.8		5.3	

Table 18. The difference in ease-of-use between qualitative and quantitative data

	Qualitative data		Quantitative data	
Ratings of Ease-of-use	Frequency	Percentage	Frequency	Percentage
1	2	7.7%	2	1.3%
2	3	11.5%	5	3.2%
3	-	-	11	7.1%
4	2	7.7%	16	10.3%
5	9	34.6%	27	17.3%
6	4	15.4%	49	31.4%
7	6	23.1%	46	29.5%
Total	26	100%	156	100%
Average ratings	4.9		5.5	

Hypothesis 4B: Survey data are perceived as better documented than data collected using some other methods.

Since census data are collected in pretty much the same way as surveys except that the census attempts to cover everybody in a population, whereas surveys study a sample in a population. I grouped surveys and census data together and then compared the perceived documentation quality with other kinds of data. Since I put survey and interview data together as one category, when doing the tests, I differentiated qualitative survey and quantitative surveys. Statistical tests showed that **documentation for survey data and census data are more sufficient (z = 3.5, p = 0.00), easier to use (z=2.3, p=0.02) and of better overall quality (z=-3.2, p=0.00) than administrative records and interview data.**

Table 19. Difference in overall quality between survey data and data collected by other methods

	Administrative records and interviews		Survey and census	
Ratings of Overall Quality	Frequency	Percentage	Frequency	Percentage
1	4	6.0%	2	0.8%
2	2	3.0%	6	2.3%
3	4	6.0%	13	5.0%
4	7	10.5%	18	7.0%
5	16	23.9%	35	13.6%
6	15	22.4%	65	25.2%
7	19	28.4%	119	46.1%
Total	67	100%	258	100%
Average ratings	5.2		5.9	

Table 20. Difference in ease-of-use between survey data and data collected by other methods

	Administrative records and interviews		Survey and census	
Ratings of Ease-of-use	Frequency	Percentage	Frequency	Percentage
1	3	4.6%	2	0.8%
2	5	7.7%	9	3.5%
3	4	6.2%	19	7.3%
4	6	9.2%	23	8.9%
5	19	29.2%	50	19.3%
6	14	21.5%	78	30.1%
7	14	21.5%	78	30.1%
Total	65	100%	259	100%
Average ratings	5.0		5.5	

Table 21. Difference in sufficiency between survey data and data collected by other methods

	Administrative records and interviews		Survey and census	
Ratings of Sufficiency	Frequency	Percentage	Frequency	Percentage
1	1	1.5%	2	0.8%
2	4	6.0%	5	1.9%
3	7	10.5%	14	5.3%
4	14	20.9%	33	12.4%
5	17	25.4%	60	22.6%
6	14	20.9%	81	30.5%
7	10	14.9%	71	26.7%
Total	67	100%	266	100%
Average ratings	4.8		5.5	

Statistics showed the connection between producer types and data types. Qualitative data are more likely to be produced by small data producers (individual researchers or small research groups) than quantitative datasets (chi^2=14.18, p = 0.00[5]). Quantitative surveys and census data are more likely to be produced by large data producers.

Data about straightforward facts are less vulnerable to the tacit knowledge problem and tend to be well documented. This happens for survey data but also can happen for other kinds of data as well. Interviewee #3 used a survey of state and local law enforcement agencies in the United States. Data include the number of employees for the agency, expenditures and average salary, equipment, etc. Interviewee #3 said that kind of information is very straightforward and very easy to use. Interviewee #2 mentioned that demographic data is very easy to use because of the straightforwardness. This kind of data are well documented not because data producers took a lot of effort in documenting data, but because those data do not need much documentation, in many cases common knowledge is enough to understand some variables.

Hypothesis 5: The stronger the users' absorptive capacity, the higher the perceived documentation quality.

Besides documentation, users' absorptive capacity is also important to use secondary data. Interviewee #5 said:

> "I think in general the documentation is often not enough. It makes a lot of assumptions that you have used that data before, can interpret what that metadata is talking about. You have to be trained to understand the code, documentation that comes with data sets. Using Geolitics[6] data as an example, I had to be very familiar with the census that was distributed, how they were distributed, the way they sample from the long form survey. I had to be familiar with that and just information provided on the CD with the database didn't give complete explanations of how the survey was conducted."

As mentioned before, users' absorptive capacity was measured by six items: 1) users'

[5] Chi-square test can be used to determine whether two variables are related to one another (independent or not).

[6] Geolitics is a company that distributes Census Bureau data.

professional status, 2) users' familiarity with the topics of data, 3) users' experience in using exactly the same data, 4) users' experience in secondary data analysis in general, 5) users' experience in collecting data and 6) users' experience in analyzing self-collected data. Users' professional status is a categorical variable but was recoded into an ordinal variable (see the details on page 42). All other five items were measured based on users' rating of how much they agree with a statement. Data showed users' familiarity with the topics of data had a highly skewed distribution. Most users are familiar with the topics of data. Users' familiarity with the same data has a bi-model distribution. This is consistent with interview findings: some users use a dataset for the first time whereas other users have used the same data many times. The highly skewed distribution happens to other items for absorptive capacity as well. (See Tables 31 and 32 in Appendix A for the value distributions of the items for absorptive capacity.)

Users' experience in collecting research data and their experience in analyzing self-collected data were highly correlated (r=0.84). Factor analysis found they load onto one factor (N=326, eigenvalue: 1.83, explained 92% of the variance). The factor loadings for the two items were both 0.96. The Cronbach's alpha for the two items is 0.91. These two items were combined into one variable: experience with primary data. The five remaining items were used to measure absorptive capacity.

Difference Between Users With High And Low Absorptive Capacity

To compare users with high and low absorptive capacity, I recoded the five measurement items of absorptive capacity. For professional status, I put professors (including full, associate and assistant professors) into one group, and I put students (including doctoral candidates, doctoral pre-candidates, masters and undergraduate students) into another group. To make sure the two groups were different, I excluded post-doctoral researchers (rank 4, right in the middle) from either group. For users' familiarity with the topics of data, users with ratings 1, 2 and 3 were put into one group; users with ratings 5, 6 and 7 were put in another group. To make sure the two groups were different, users with rating 4 were excluded from either group. The same grouping procedure was applied to the other three items for absorptive capacity as well. I then compared each pair of groups and found significant difference between all the groups. Below are the details.

Professors (including full, associate and assistant professors) perceive the documentation they use as more sufficient (z=-3, p=0.00), easier to use (z=-1.8, p=0.07) and better overall (z=-2.3, p=0.02) than students (including doctoral candidates, doctoral pre-candidates, masters and undergraduate students). (See Tables 33, 34 and 35 in Appendix A for more detailed differences between students and professors. Users familiar with the topics of the data (ratings 5, 6, 7) perceive the documentation they use as more sufficient (z=-3.4, p=0.00), easier to use (z=-2.6, p=0.01) and better overall (z=-3.7, p=0.00) than users not familiar with the topics of data (ratings 1, 2, 3). (See Tables 36, 37 and 38 in Appendix A for more details.) Users experienced with the same data (ratings 5, 6, 7) perceive the documentation they use as more sufficient (z=-5.4, p=0.00), easier to use (z=-4.0, p=0.00) and better overall (z=-3.0, p=0.00) than users not experienced with the same data (ratings 1, 2, 3). (See Tables 39, 40 and 41 for more details.) Users experienced in secondary data analysis (ratings 5, 6and 7) perceive the documentation they use as more sufficient (z=-2.9, p=0.00), easier to use (z=-2.6, p=0.01) and better overall (z=-1.7, p=0.08) than users not experienced with the same data (ratings 1, 2, 3). (See Tables 42, 43 and 44 for more details.) Users experienced in collecting and analyzing self-collected data (ratings 5, 6 and 7) perceive the documentation they use as more sufficient (z=-2.5, p=0.01) than users not experienced in collecting and analyzing self-collected data (ratings 1, 2 and 3). (See Tables 45, 46 and 47 for more details.)

Comparison Across Different Professional Ranks

Professional status was recoded into an ordinal variable with 8 levels. However, since the sample sizes for doctoral pre-candidates and undergraduates were too small to have statistical power, doctoral pre-candidates, master students and undergraduates were put into one group. Then I compared the perceived documentation quality of the remaining 6 professional ranks. Significant differences across professional ranks were reported below.

Full professors perceive the documentation they use as easier to use than associate professors (z=-1.7, p=0.09). Full professors perceive the documentation they use as more sufficient (z=-2.5, p=0.01), easier to use (z=-3.1, p=0.00) and better overall (z=-1.7 p=0.08) than post-doctoral researchers. Full professors perceive the documentation they use as more sufficient (z=-2.4 p=0.01) and easier to use (z=-1.8, p=0.07) than doctoral candidates. Full professors perceive the documentation they use as more sufficient (z=-2.9, p=0.00), easier to use (z=-2.4, p=0.02) and better overall (z=-1.8, p=0.07) than doctoral pre-candidates, master students and undergraduate students.

Associate professors perceive the documentation they use as easier to use (z=-1.9, p=0.05) than post-doctoral researchers. Associate professors perceive the documentation they use as more sufficient (z=-1.8, z=0.07) than doctoral pre-candidates, masters and undergraduates.

Assistant professors perceive the documentation they use as more sufficient (z=-2.2, p=0.03) and easier to use (z=-2.04, p=0.04) than post-doctoral researchers. Assistant professors perceive the documentation they use as more sufficient (z=-2.0, p=0.05) than doctoral candidates. Assistant professors perceive the documentation they use as more sufficient (z=-2.5, p=0.01) than doctoral pre-candidates, masters and undergraduates. Post-doctoral researchers perceive the documentation they use as harder to use (z=1.8, p=0.06) than doctoral candidates. This is the only exception that users with higher professional ranks perceive the documentation they use as worse than users with lower professional ranks. All other results have shown that users with higher professional ranks perceive the documentation they use as better or at least the same as users with lower professional ranks.

Table 22. Difference in perceived documentation quality across professional ranks

Professional Rank	Sufficiency	Ease-of-use	Overall quality	
Full professor	=	>***	=	Associate professors
Full professor	=	=	=	Assistant professors
Full professor	>*	>*	>***	Post-docs
Full professor	>*	>***	=	Doctoral candidates
Full professor	>*	>**	>***	Pre-candidates, masters and undergraduates
Associate professor	=	=	=	Assistant professors
Associate professor	=	>**	=	Post-docs
Associate professor	=	=	=	Doctoral candidates
Associate professor	>***	=	=	Pre-candidates, masters and undergraduates
Assistant professor	>**	>**	=	Post-docs
Assistant professor	>**	=	=	Doctoral candidates
Assistant professor	>*	=	=	Pre-candidates, masters and undergraduates
Post-doc	=	<***	=	Doctoral candidates
Post-doc	=	=	=	Pre-candidates, masters and undergraduates
Doctoral candidate	=	=	=	Pre-candidates, masters and undergraduates

*: p<0.01, **:p<0.05; ***:p<0.10

Partial Correlations Between Absorptive Capacity And Perceived Documentation Quality

The results above have shown that users with high absorptive capacity differ significantly from users with low absorptive capacity in their perception of documentation quality. However, it is not clear whether the difference was caused by absorptive capacity, or was caused by other confounding factors. To explore this further, I looked at the partial correlations between absorptive capacity and perceived documentation quality while controlling other impacting factors.

As shown earlier, four possible impacting factors were identified and the effects of three of them have been tested. However, those three factors are associated with each other. Data produced for sharing are more likely to be distributed by intermediaries (chi^2=3.16, p=0.08). Data produced for sharing are less vulnerable to the tacit knowledge problem: qualitative data are less likely to be produced for sharing (chi^2=13,08, p=0.00) than quantitative data; survey and census data are more likely to be produced for sharing (chi^2=111.61, p=0.00) than data collected by other methods (administrative records and interviews). According to the statisticians at the Center for Statistical Consultation and Research (CSCAR), it is not appropriate to control variables that are significantly associated with each other at the same time. Therefore only data producers' incentive (e.g. for sharing or self-use) was controlled in checking the partial correlations between absorptive capacity and perceived documentation quality.

A new variable "absorptive capacity" was created as the average of all the five items: professional status, familiarity with the topics of data, experience using the same data, experience in secondary data analysis and experience with primary data. Table 23 shows the partial correlation of each of the five items while controlling other items and data producers' incentive. Table 24 shows the partial correlations of absorptive capacity and perceived documentation quality while controlling for data producer's incentive. Consistent with the support of Hypothesis 1, whether data is produced for sharing or not is a significant predictor of all aspects of perceived documentation quality even when absorptive capacity is controlled. Users' familiarity with the topics of data is the strongest predictor of perceived documentation quality among all the five measurement items of absorptive capacity. It has significant correlations with all aspects of perceived documentation quality. The next strongest predictor is users' experience with the same data. It has significant correlations with perceived sufficiency and perceived ease-of-use, but not with overall quality. Users' experience in collecting and analyzing primary data and their experience in analyzing secondary data have significant correlations with perceived sufficiency.

Table 23. Partial correlations between perceived documentation quality and the five measurement items for absorptive capacity

Measures for Absorptive Capacity	Sufficiency	Ease-of-use	Overall quality
Professional status	-	-	-
Experience in secondary data analysis	0.12***	-	-
Experience in collecting and analyzing primary data	0.16**	-	-
Experience in analyzing the same data	0.18*	0.17*	-
Familiarity with topics of data	0.26*	0.26*	0.32*
Produced for sharing or not	0.20*	0.13**	0.23*

*: p<0.01, **: p<0.05, ***: p<0.10

Table 24. Partial correlations between perceived documentation quality and absorptive capacity

	Sufficiency	Ease-of-use	Overall quality
Absorptive capacity	0.32*	0.23*	0.19*
Produced for sharing or not	0.22*	0.14**	0.23*

*: p<0.01 level, **:p<0.05

The above results tell us that absorptive capacity has significant effects on perceived documentation quality. However, there are two possible causes of these effects. First, users with stronger absorptive capacity know how to choose data with high quality documentation. The documentation they use is perceived as better independent of users' absorptive capacity. Second, documentation quality is randomly distributed among users with different levels of absorptive capacity. The fact that some users perceive the documentation as better is because of their stronger absorptive capacity. To put it another way, for the same documentation, users with stronger absorptive capacity perceive it as better than users with lower absorptive capacity. To tease out these two different reasons, I need to look at the correlation between each aspect of absorptive capacity and perceived documentation quality when documentation quality is controlled. In other words, I need to see how people with different levels of absorptive capacity perceive identical documentation differently. But not enough data were collected in this study to conduct that test. It is true that some data were used by more than 10 respondents in this study. However, to identify identical data sets, I would need to control both the titles of data sets and also the distributors of data sets. As shown before, data distributed by intermediaries and producers have different perceived quality. That made the qualified sample too small to have statistical power. Future study is needed to tease out the two reasons.

Validity Of The Measures For Absorptive Capacity

Based on the kinds of information that users sought during their secondary data use, I identified three components of users' absorptive capacity: knowledge about the data, background knowledge to understand and interpret data, and data analysis skills. Below are some examples of the three types of knowledge.

- Knowledge about data: what is the response rate and sampling frame for a particular survey, or how are the missing responses treated.

- Background knowledge: disciplinary consensus on how to use common types of data, how to determine whether or not to weigh variables from samples, which variable best captures certain concepts, how to interpret frequently occurring variables, and how to handle specific measurement issues.
- Data analysis skills: how to convert hierarchical data files into appropriate rectangular files, how to construct new derived variables, how to use a statistical method such as linear regression, etc.

According to existing literature (Zimmerman, 2003) and the interview findings above, background knowledge and data analysis skills are accumulated through academic training and professional experience. Unlike documentation, which is external knowledge, background knowledge and data analysis skills are internalized as part of secondary users' knowledgebase. Some knowledge about data is provided by documentation of data. Knowledge from documentation can be internalized and become part of users' absorptive capacity. In those cases, users would be less reliant on documentation. This is most dramatic when a user has used the same dataset or the same data series many times.

Based on the findings about the components of absorptive capacity, I analyzed the validity of the five measurement items of absorptive capacity. It seems that the item "familiarity with the topics of data" directly measures users' background knowledge related to the particular data being used. The item "experience using the same data" does not directly measure users' knowledge about the particular data being used, but is a strong predictor of users' familiarity with data. It is quite reasonable to believe that if a user has used a particular dataset before, he/she has gained some knowledge about the data. The variable "experience in secondary data analysis" and the variable "experience analyzing primary data" does not directly measure users' data analysis skills. According to Interviewee #5, data analysis experience builds up data analysis skills. Experience analyzing primary and secondary data should be a predictor of data analysis skills in general. But it does not necessarily mean a user has no difficulty in analyzing a particular dataset if he/she has experience in analyzing other datasets. Interviewee #7 is a sociologist with a lot of experience in secondary data analysis. But she encountered difficulties the first time she analyzed longitudinal dataset using a new statistical method that she had never used before. Therefore experience in secondary and primary data analysis is not a strong predictor of data analysis skills for a particular use case. Users' professional status and experiences in collecting primary data do not directly measure any aspect of absorptive capacity and do not look like strong predictors of any aspect of absorptive capacity for a particular dataset either. Therefore the three items (professional status, experience with primary data, and experience with secondary data analysis) may not be used as indicators of users' absorptive capacity for particular datasets.

Users' familiarity with the topics of data use is about the knowledge distance between the data and data users in terms of the intellectual content. In addition to the knowledge distance in terms of intellectual content, interview data revealed the knowledge distance in terms of data analysis tools and methods. Interviewee #5 mentioned the differences in the methodology and tools in analyzing data from different disciplines:

> "The primary differences involve the methodology and software that are best suited to the different disciplines. In economics, I primarily work with time series data, or longitudinal data (using SAS, SPSS, Stata, EViews, etc.), while in geography the data are spatial and tend to be cross-sectional (requiring the use of Relational Databases and Geographic Information Systems)."

In a future study, measures for absorptive capacity need to be re-created based on these findings.

Summary And Discussion

In this section, the effects of the four impacting factors have been confirmed. Data produced for sharing are better documented than data produced for self-use. Intermediaries are effective in improving documentation quality for data produced for sharing. Data produced for sharing and distributed by intermediaries are better documented than data produced for sharing and distributed by data producers. Data less vulnerable to the tacit knowledge are better documented than data more vulnerable to the tacit knowledge. More specifically, quantitative data are better documented than qualitative data. Quantitative survey and census data are better documented than data collected by qualitative interviews and administrative records. Data about straightforward facts tend to be better documented. Users with stronger absorptive capacity perceive the documentation they use as better than users with weaker absorptive capacity.

Based on the effects of these impacting factors, causes of perceived inadequacy of documentation can be grouped into three categories: inherent inadequacies of documentation, poor documentation and possibly users' low absorptive capacity. Inherent inadequacies are caused by tacit knowledge and communication reduction. This problem reveals the inherent limitation of documentation as a knowledge transfer channel. The solution to this problem is less about improving the quality of documentation and more about providing other knowledge transfer channels. Data are poorly documented either because data producers are not motivated or unable document data well for secondary use. This problem could be solved or alleviated by providing appropriate incentives or instruction to data producers, or involving intermediaries to improve data and documenting quality. Users with lower absorptive capacity perceived the documentation they use as less well documented than users with higher absorptive capacity. In this case, users themselves, data producers, or intermediaries should take effort to improve users' absorptive capacity.

The Effect Of Perceived Documentation Quality On Users' Incentive To Use Secondary Data

The benefit of using secondary data versus collecting primary data is obvious. As Interviewee #3 said: "The data are there, they are already collected, it saves a lot of work, a lot of expense." However, users are not always willing to use secondary data because there are also costs associated with secondary data use. Inadequate documentation increases use cost and may turn users away in some situations. However, users' incentives to use secondary data mostly depend on how well the data fit their information needs rather than documentation quality.

Sometimes a piece of information needed by a secondary user does not exist in the data and documentation. There are two possible reasons for this. 1) Data producers did not collect that information, in other words, that information does not exist at all. In that case, users would not use the data no matter how well documented the dataset is. Interviewee #3 had a research question on his mind and he heard about a dataset that might answer his question. After he explored the data for a while, he found that data set actually did not answer his research question even though it was a nice dataset that was well documented. So he gave up using it. Interviewee #5 expressed similar views: even though many questions were asked about a certain topic in collecting a dataset, there may be one question left unasked and that happened to be just what a secondary user needs. 2) The information could exist but is not documented. In other words, the data is not well documented. Users may take effort to search for that information. But if the cost of searching is greater than the benefit they could possibly get from using data, users still would give up using the data. Interviewee #5 gave up using a geographic dataset because the projection system [7] was not recorded, which should be solved by better documentation. The cost of finding the projection system may not be that high. However, that dataset was not important for her; it is only something that she was experimenting with, and there were other datasets with better documentation available. So in this case, even though the cost may not be high, the benefit is not high enough to justify the cost. So she gave up using that dataset as well.

In many cases, a dataset does contain relevant information, but not exactly what the users need. Users may need to compromise their information needs if they use the data. When talking about the secondary use of data produced by other individual researchers, interviewee #2 said she prefers to collect her own data even if the secondary data are well documented, because she wants the data to be tailored specifically to the research question that she wants to explore. Interviewee #3 gave an example about the measurement of wealth: people want to look at the relationship between wealth and

[7] The projection system is geographic coordinate system. There are dozens of different geographic coordinate system. If it is not well documented the type of coordinate system is unknown.

crime. But the secondary data they obtained do not contain a perfect measure of wealth. A relevant variable in the data is whether people own or tent home. Some users just use that variable to measure wealth because that is the only thing they have. A similar thing happened for interviewee #7 when she tried to use secondary data to measure peer delinquency. Even if data about the same concepts were collected in secondary data, the operationalization of those concepts may be slightly different from what the secondary user wants. Interviewee #6 pointed out that leads to another cost of using secondary data: learning how someone else decided how to ask a question or coded a variable in a certain way.

Consistent with existing literature (Zimmerman, 2003), uncertainty was found to be very common in using secondary data. Inadequate documentation undoubtedly causes uncertainty. But perfect documentation would not always eliminate uncertainty. It is consistent among interviewees that even if data are very well documented, there are still often uncertainties in secondary data use. Sometimes that uncertainty doesn't affect data use. As interviewee #6 said:

> "It is not your own data. You always feel you don't understand it as well as if it was your own data. I don't understand it fully, but that doesn't mean I don't understand it well enough to use it. "

Some uncertainty could be partially solved after seeking outside information. When uncertainty cannot be solved, users would need to decide whether to tolerate the uncertainty or give up using data. Interview #5 described how she dealt with uncertainties in using secondary data: if she is really not sure about something and can't get help about it, she won't proceed on it. Sometimes she had to rely on her best interpretation of what to do and leave it to peer review later.

If a dataset is important and there is no way for researchers to collect the same data themselves, users would still try to use the dataset even if the data is not well documented. In this scenario, the cost of using data may be high, but the benefit of using data is high enough to justify the use. This happens mostly for administrative records produced for self-use. Users may be reluctant to use the data, but they have to use them because there is no easier way to get the information. Interviewee #1 got a database about Texas alcohol and beverage control. There were so many errors and missing data that he postponed in using it, but he did not give up using it. Interviewee #5 still used a dataset about American religion even though there are problems with the data and documentation, because it is the best that is available.

Many researchers rely almost exclusively on large-scale data for their research. For them, using secondary data produced for sharing is a must and routine because they cannot collect data themselves on such a large scale. Documentation of data produced for sharing is good in general but not always perfect. Some users reported that the documentation of data produced for sharing is massive and hard to use. Even with these problems, no users said they would not use the data solely because documentation is not ideal. In fact, users expressed appreciation, understanding, and forgiveness for problems with documentation for datasets that are produced for sharing. One user said: “This dataset is extremely extensive and complex by nature. The documentation was very good given the data complexity.” Some users of this kind of data are willing to take on the additional burdens associated with using large complex datasets. They think it is natural that documentation does not provide everything. It is part of their research process to interact with other researchers for secondary data use. Some users understand that producing very high quality documentation is time consuming, and may mean that release of data would be delayed if the documentation had to be ideal, especially for data that are collected every couple of years. There is a choice on how to allocate limited resources. They prefer data to be produced as efficiently as possible, instead of waiting longer for the most complete and fantastic metadata. Interviewee #5, who is a heavy user of large scale survey and census data, appreciates data producers so much that she has decided to answer every academic survey request.

In the case that users decide to use secondary data, information seeking is very often necessary to supplement documentation, reduce uncertainty and improve users’ absorptive capacity.

How Do Users Overcome Inadequate Documentation?

A simple answer to this research question is: to seek information not provided in documentation so that users' information needs can be fulfilled. Information seeking is important for almost every stage of secondary data use, starting from the process of searching and obtaining data. When asked about the challenges in using secondary data, several interviewees answered the difficulties in knowing where the data are and how to get them. This study focuses on users' information seeking after they obtain and decide to use particular datasets.

Users need external documentation of data and an internal absorptive capacity to analyze secondary data. Consequently, either inadequate documentation or an underdeveloped absorptive capacity would motivate them to seek outside information to fulfill their needs. In this dissertation, I use the word "outside information" to refer to knowledge and information necessary for secondary use but missing from either the documentation or the users' absorptive capacity.

Documentation can be adequate enough so that users with sufficient absorptive capacity can use data solely based on documentation. Actually 19% of the respondents used secondary data based only on documentation. However, seeking outside information is often necessary. 161 (46%) of surveyed users obtained outside information because the documentation did not contain the information they needed (not sufficient). In addition, hard-to-use documentation also turns users to other information sources. As one user said:

> "The code book contains the definitions of variables, but sometimes I think it is easier to pick up the phone. I got the codebook, but sometimes the variables are not clear to me. They do have a user manual, as well as statistical manual. Sometimes if I know someone who might know the answer, it is easier for me to pick up the phone."

44 (12.5%) of surveyed users sought outside information because the documentation was hard to use. 110 (31.4%) of the survey respondents obtained outside information because other information sources or channels are immediately accessible. This is useful for understanding the following phenomenon: even though data produced for sharing are

better documented than data produced for self-use (Niu & Hedstrom, 2009), users of data produced for sharing are just as likely to seek outside information as users of data produced for self-use. The reasons could be: producers who produce data for sharing tend to provide user assistance and make outside information more readily accessible. A few users sought outside information because they detected some errors in data and documentation (accuracy issues). People who did not seek outside information rate documentation as more sufficient (z=5.7, p=0.00), easier to use (z=4.2, p=0.00) and of better overall quality (z=2.6, p=0.01) than users who did.

About one-quarter of the respondents collaborated with data producers in using secondary data. I found that people who collaborated with producers perceive documentation as less sufficient (z=-2.3, p= 0.02) than users who did not collaborate. At the same time, users of data produced for self-use are more likely to collaborate with data producers than users of data produced for sharing (chi^2 = 6.63, p = 0.01). Qualitative data users are more likely to collaborate with data producers than quantitative data users (chi^2= 5.45, p = 0.02). Therefore it may not be that collaboration with data producers causes users to perceive documentation as less sufficient, but that users of less sufficiently documented data tend to collaborate with data producers in order to use the data. Users who obtained data directly from data producers are also more likely to collaborate with data producers than users who obtained data from intermediaries including data archives and data centers (chi^2=9.5, p=0.00).

Some information seeking is driven by the need to build or increase absorptive capacity. Users with lower professional status[8] are more likely to seek outside information (z=2.5, p=0.01). This is consistent with interview findings: students often get data analysis help from their advisors. As mentioned earlier, absorptive capacity is accumulated through academic training and professional experience. Therefore most of it is acquired before users start up using a particular dataset. As Interviewee #5 said: she would not be at the point of getting and analyzing data if she is not familiar with some terminology in the

[8] Professional status means whether the user is a full professor, associate professor, assistant professor, post-doctoral researcher, doctoral student, etc.

data. But often the case is that users did not expect something in a dataset during the process of analyzing it. For example, interviewee #6 used court records. Even though her field is criminal justice, sentencing and policing, she still needs to learn something about the court process to use the data. To analyze a dataset using a new statistical skill, interviewee #7 took a summer course offered by a data archive. Interviewee #5 received a lot of help from her advisor about how to read and interpret the data, and to select variables and conduct analysis while she analyze secondary data the first time. Here are more examples from the survey:

> "I did not receive information (about the data) outside of the documentation. What I received was additional information on how others used/interpreted the same data. This provided me with a deeper understanding of the data which I think was a benefit when I used the same data sets." "The information was not on the datasets itself but rather on operation of derived variables in previous studies. For instance, how do you handle negative incomes and the like." "It was an analytic question more than a data-specific question."

Besides adequate documentation and low absorptive capacity, there are some social and psychological reasons that users seek outside knowledge. Consider the following reasons that users sought outside information.

> "Working with others helps to bring the data alive." "Just want to get as much information as possible" "It was nice to talk to someone about the data instead of just read about it." " It is useful to talk with other secondary data users who are more knowledgeable about the data." "Website provided additional information that was useful for my analysis, but I didn't need to use the website to use the data."

There are also 41 users (11.6%) who obtained outside information not because they actively sought it, but because they happened to encounter that information.

These findings contradict an important assumption of this study: when documentation is adequate, users do not need outside information to use the data. The above findings offer new insights into several things. When documentation is adequate for some users, users with lower absorptive capacity may perceive it as inadequate. Users have some social and psychological reasons for obtaining outside information even though documentation is adequate. Existing literature mentioned that users often obtain outside information because documentation is hard to use even though sufficient (Sieber, 1991). While that is true in many circumstances, this study shows that users may still seek outside information even though the documentation has no problems. People like to get information through socializing and talking rather than reading alone. These findings shows the importance of providing other knowledge transfer channels and building communities of data users to help overcome problems with documentation quality.

The fact that insufficient documentation is not the only reason that users seek outside information shows that users' outside information seeking is not a good indicator of documentation sufficiency. This also explains why the correlation between some indicators of sufficiency is moderate. Measurements of sufficiency need to be improved in future study.

Sources And Channels Where Users Seek Outside Information

As mentioned earlier, I categorized knowledge transfer channels into three types: the use of documents, interactive conversations and situated learning. Only explicit knowledge can be transferred through documents. Knowledge transferred through the interactive conversations channel is primarily explicit knowledge that is verbalized and not formally documented. Tacit knowledge that is very hard to articulate, but it can be transferred through situated learning such as visiting or working together. The three channels are used as rationale for analyzing the sources and channels where users obtained outside information.

Table 25 is a list of outside information sources and the percentage of users who used each source for outside information. Please be informed that one user may seek outside information from several sources.

Table 25. Sources for outside information (N=353)

Previously written articles using the dataset	47%	Documents (64%)
Websites of data producers	34%	
Websites of data archives	17%	
Data producers	41%	People (68%)
Other secondary users	40%	
Data archivists	13%	
Workshop	8 %	

Besides those main sources, other information sources include: using related datasets to check the integrity of the data or for other reasons, publications based on similar data collected by different researchers, outside sources of scales used in the data set, alternative publications with similar information, newsletters and mailing lists for users of the same data, relevant newspaper, etc. 64% of the respondents obtained outside information from various kinds of documents. Based on the rationale above, that kind of knowledge is explicit knowledge that can be incorporated into documentation, or at least pointed to from documentation. Actually, the abundance of these related documents is one reason why well known and heavily used datasets are easier to use. Interviewee #5 said: “I think that well-known and used data tend to be easier to use because there are more articles and working papers to draw on as references.” Previous publications based on the same data are the documents most frequently used. This is consistent with interview findings. Several interviewees directly or indirectly mentioned the importance of previous publications. Interviewee #3 suggested examples of the uses of the data be included in documentation: “It might be helpful if there were examples of the uses of the data. You can use this data to do this, or do that. That gives another bit of information to the user. ” Previous publications are often examples of the uses of data.

More than 60 percent of users sought outside information from other people. Some of those people work closely with secondary data users, such as mentors, advisors and colleagues. Some are strangers, such as other users of the same data on a Listserve, other data users or data producers found through Internet search, data producers who left contact information in documentation, and data archivists where users obtained the data. Users who obtained data from data producers are more likely to seek outside information from data producers (chi^2=6.87, p=0.01) and websites of data producers (chi^2=5.59, p=0.02). Among 239 users who obtained information from people, 80% obtained that information through email or telephone, 55% obtained that information through face-to-face conversations, 31.4% obtained that information by working together with other people. 18.4% used all of the three channels. About half of the users (49%) used at least two channels. 36% only used email or telephone, 11% only used face-to-face conversations, 3% obtained knowledge only by working together with other people.

For users who obtained outside information only through one channel, Table 26 showed the distribution of reasons why users sought outside information. The five letters represent five reasons why users obtained outside information. Please be informed some users sought information from a channel for more than one reason. That is why the sum of all five reasons is greater than 1.

Table 26. Reasons for obtaining outside information (N=353)

Reasons	Channels		
	Email/ telephone	Face-to-face	Work together
A. Documentation does not contain information they need.	66%	31%	29%
B. Documentation is hard to use.	16%	8%	0%
C. Other information sources and channels are immediately accessible.	28%	58%	57%
D. They happen to encounter that information.	5%	19%	29%
E. That information is tacit knowledge that is hard to document..	18%	19%	43%

A (sufficiency) and B (ease-of-use) are related to the adequacy of documentation, C (outside information immediately accessible) and D (happen to encounter outside information) are related to the convenience of information sources and channels. Table 27 shows the percentages of respondents who sought outside information because of inadequate documentation (either A or B), convenient outside information (either C or D) and the tacit knowledge problem (E). In the table, percentages of C or D are smaller than the sum of C and D. That is because some users sought information through one channel for both C and D. For example, 57% of users sought information only through working together because outside information were readily accessible (C), which can be interpreted this way: they happened to work together with people who provided help, not because they have to work together with other people to obtain that information. 29% of users sought outside information only through working together because they happened to encounter that information (D). Instead of 86%, only 71% of users sought information only through working together because of either C or D, because there are 15% of users sought information because of both C and D. This overlap is quite reasonable because if people happen to work together, it is easier for them to encounter information passively without actively seeking information.

Table 27. Combined reasons for obtaining outside information (N=353)

Combined Reasons	Channels		
	Email/ telephone	Face-to-face	Work together
Inadequate documentation (either A or B)	72%	35%	29%
Convenient outside information (either C or D)	31%	65%	71%
E	18%	19%	43%

We can see several patterns from Table 27. First, people are more likely to obtain outside information through email and telephone when documentation is inadequate. Second, people are more likely to obtain outside information through face-to-face or working together when those channels are convenient. Third, people are more likely to obtain hard-to-document tacit knowledge through working together**.** This third pattern is consistent with our rationale that tacit knowledge is more suitable to be transferred through situated learning than through documents or interactive conversations. Among 21 users who sought outside information only because of the tacit knowledge problem, 18 sought that information from people (17 email and telephone, 10 face-to-face conversations, three through working together); 14 obtained outside information from various documents (websites and publications based on the datasets). Seven obtained information solely from people, three obtained that information solely from documents. A higher percentage of qualitative data users reported tacit knowledge problems than quantitative data users (chi^2=3.3, p=0.07). This is consistent with the proposition that qualitative data are more vulnerable to tacit knowledge problems. As we know, only knowledge articulable can be transferred through documents. Therefore the fact that tacit knowledge is transferred also through documents confirmed that some people regard tacit knowledge more broadly than knowledge technically hard to articulate. One type of tacit knowledge found in this study belongs to the category defined by Collins (2001): missing knowledge caused by a mismatch between data producers' and users' concerns. As one user said:

> "while you are using existing data, most of the time somebody collected it for a different reason. The failure that I had with the sentencing data was that whoever collected that data, for whatever reason didn't need to know where the offender came from. So they didn't record it."

This study also revealed a new category of tacit knowledge that tends to be missing from documentation: Informal knowledge. One user said: "People don't document why some of the numbers are funny, things that went wrong in the survey, etc." The tacit knowledge problem and the fact that users seek information from multiple sources and channels reinforce the importance of providing other knowledge transfer channels besides documentation.

Summary

In this chapter, the research questions were answered based on the combined analysis of survey and interview data. Before answering the research questions, the reliability and validity of the DEM were tested. I found the DEM was reliable and valid in general with several exceptions. Hypothesis tests proved that effects of the proposed four impacting factors of perceived documentation quality. Inadequate documentation increases use cost and may turn users away in some situations. However, users' incentives to use secondary data mostly depend on how well the data fit their information needs rather than documentation quality. Outside information seeking is very common in secondary data use. Users seek information because of inadequate documentation (insufficient, hard to use, inaccurate), inherent limitations of documentation, users' low absorptive capacity, and for convenience and social and psychological reasons. In seeking outside information, users tend to consult multiple sources and use several communication channels. Publications based on the same data are the most frequently used sources of outside information. Email or telephone is the primary channel through which users seek outside information from people.

CHAPTER 5

CONCLUSIONS

Perceived documentation quality is affected by data producers' incentive and capacity to document data, secondary users' absorptive capacity in using data, the existence of intermediaries to improve documentation quality, and data's vulnerability to tacit knowledge problem. Data produced for sharing are better documented than data produced for self-use. Users with stronger absorptive capacity tend to perceive the documentation they use as slightly better than users with weaker absorptive capacity. Intermediaries such as data archives have been effective in improving documentation quality of data produced for sharing. Data less vulnerable to tacit knowledge problem, such as quantitative survey data and data about straightforward facts, are perceived as better documented than data more vulnerable to tacit knowledge problem, such as qualitative data.

Inadequate documentation increases use cost and may turn users away in some situations. However, users' incentives to use secondary data mostly depend on how well the data fit their information needs rather than documentation quality. Users would not use a dataset if it doesn't answer their research questions, no matter how well documented it is. When data and documentation do not fit users' needs perfectly, users need to decide whether to compromise their information needs or give up using the data. Their decision-making is not necessarily changed by documentation quality. With inadequate documentation, users need to seek outside information to supplement documentation and reduce uncertainty. Their decision to use or not depends on how much they can benefit from using the data, the cost of overcoming inadequate documentation and the potential cost to them to collect the same data. As a result, many users do not want to use small secondary datasets because the potential cost of collecting the same data is not greater than the costs of using secondary data, which are caused by uncertainties, information seeking and the compromization of information needs. On the other hand, even though many administrative records are regarded as messy and poorly documented, users still often choose to use those data because there is no way they can collect the same data.

Some users can use some data based solely on the documentation of data. But more often users need to seek information not provided in documentation to use secondary data. Inadequate documentation (insufficient, hard-to-use, inaccurate) is not the only cause of outside information seeking. Users of good documentation still seek or obtain outside information because of the tacit knowledge problem, because their absorptive capacity is not ready to use the data, because they feel psychologically more comfortable in obtaining information through socializing than through reading documentation alone, or because they encounter useful information without actively seeking. In seeking outside information, users tend to consult multiple sources and channels. Publications based on the same data are the most frequently used sources of outside information. Email or telephone is the primary channel through which users seek outside information from people.

Implications For Strategies To Help Secondary Data Use

An assumption of this study is: with adequate documentation, users do not need to seek outside information to use secondary data. In line with that assumption, the strategy to help secondary data users is to find ways to make documentation adequate enough so that users can use data solely based on documentation. Whereas findings from this study tell us improving documentation quality is not always the right way to help secondary data users.

Documentation of data produced for sharing tends to be good enough for secondary use. Even though not perfect, secondary users would prefer the data released sooner than waiting longer for better documentation, and they are willing to take the effort to overcome imperfect documentation. Other knowledge transfer channels are often necessary for secondary data use even the data is well documented.

Due to the tacit knowledge problem and communication reduction, some kinds of documentation are inherently insufficient, there is no way or very costly to improve its adequacy. In this case, providing other knowledge transfer channels between data producers and secondary users, and among secondary users is more appropriate than improving documentation quality.

Users' with stronger absorptive capacity perceive the documentation they use slightly better than users with lower absorptive capacity. Therefore training data users also would help secondary data users.

Data produced for self-use tend to be poorly documented. Providing instruction and incentives to data producers is the right way to improve these kinds of documentation.

Limitation Of This Study

Since no existing metrics for perceived documentation quality and users' absorptive capacity were found before conducting this study, the evaluation metrics for them were created based on literature review and the preliminary analysis of interview data. This is both contribution and limitation of this study. Two indicators for the construct ease-of-use (easy to find information from the documentation; the content of the documentation is clear and understandable) are highly reliable and valid. They can be re-used by other researchers. The metrics for sufficiency and absorptive capacity are not strongly reliable but were used to answer the research questions. That might affect the statistical results and conclusions.

Data and documentation are very closely related. When I interviewed secondary data users about documentation quality, sometimes they stray away to talk about the quality of data. For many users, documentation is part of data. I chose to study only documentation quality because of convenience. I wanted to make this project smaller and easier to manage. Documentation is the metadata of social science data. As a metadata person, it is a natural tendency for me to focus on metadata. The consequence of this choice is an incomplete study for problems in secondary data use.

Even though claimed to study secondary users of social science data, I mainly studied users of survey / interview data, census data and administrative records. Very few users of experimental data and observation data were collected in the sample. An important reason is those kinds of data are not as commonly used in practice. It is not clear whether findings from this study apply to experiment data and observation data.

Future Work

As mentioned before, data and documentation are very closely related. A study of the quality of both data and documentation would reveal a more complete picture of problems in secondary data use, and be more informative about how to help secondary data users. Even though experimental data and observation data are not commonly used as survey data and administrative records, a purposive sampling strategy might help collect enough users of those kinds of data, and make this study more complete. Data captured by various virtual communities, such as online stores, online communities, gaming environments, etc., are more and more commonly used for social science research. However, as far as I know, not much research has been conducted on the sharing, management and use of those data. Data archives do not preserve and disseminate those kinds of data yet. It would be interesting to look into the secondary use of those kinds of data as well.

This study has shown that users with stronger absorptive capacity perceive the documentation they use as better than users with lower absorptive capacity. As mentioned before, there are two possible causes of that difference. First, users with stronger absorptive capacity know how to choose data with high quality documentation. The documentation they use is perceived as better independent of users' absorptive capacity. Second, documentation quality is randomly distributed among users with different levels of absorptive capacity. The fact that some users perceive the documentation better is because of their stronger absorptive capacity. To tease out these two different reasons, an experimental study could be conducted to look at how people with different levels of absorptive capacity perceived the same documentation differently. To conduct that study, better measurements of absorptive capacity needs to be created based on the results from this study.

An impacting factor of documentation quality is the incentive of data producers. The natural solution to that problem is to provide incentives to data producers. A possible kind of incentive mechanism is to make data producers liable for messy data and poor documentation. Whether that incentive mechanism is a good choice for the society depends on two conditions: first, the negotiation cost between data producers and secondary users; second, the cost of data producers to provide adequate documentation and the cost of users to overcome inadequate documentation. According to the Coase rule (Frank, 2007, p. 539 and p. 543), when the parties affected by externalities[1] can negotiate costlessly with one another, an efficient outcome results no matter how the law assigns responsibility for damages. When there is a negotiation cost, efficient laws and social institutions are the ones that place the burden of adjustment to externalities on those who can accomplish it with least cost. Messy data and poor documentation provided by data producers is a kind of externality to secondary data users. Data producers can solve the problem by taking time and effort to improve the quality of data and documentation. Secondary users can overcome inadequate documentation by seeking outside information, making compromises and tolerate uncertainties, etc. If the negotiation between the two parties is cost free, an efficient outcome results no matter how the law assigns responsibility for messy data and poor documentation. In other words, it is not necessary to make data producers liable. If the negotiation cost between the two parties cannot be ignored, to achieve higher efficiency for the society, whether we should make data producers liable depends on which party can solve the problem with lower cost, which may vary with different data producers and users. It may not be cost efficient for a government agency to improve documentation only for a few secondary users, but cost efficient for a very large number of users. If a researcher requests data from another researcher, it is not clear which party can solve the externality with lower cost. A future study can be conducted to investigate the negotiation cost between different types of data producers and data users. This would help decide whether it is more efficient for the society to make data producers liable for messy data and poor documentation.

APPENDICES

APPENDIX A

Extra Tables

Table 27. Value distributions of the 19 items for completeness

		1	2	3	4	5	6	7	N/A
Title	Freq.	5	9	1	8	13	27	288	15
	Percent.	1.4%	2.5%	0.3%	2.2%	3.6%	7.4%	78.7%	4.1%
Principal investigator	Freq.	13	15	5	13	28	37	189	64
	Percent.	3.6%	4.1%	1.4%	3.6%	7.7%	10.2%	51.9%	17.6%
Time period	Freq.	10	5	2	10	20	34	280	6
	Percent.	2.7%	1.4%	0.5%	2.7%	5.5%	9.3%	76.3%	1.6%
Geographic location	Freq.	14	12	7	13	26	52	228	14
	Percent.	3.8%	3.3%	1.9%	3.6	7.1%	14.2%	62.3%	3.8%
Funding agency	Freq.	9	7	13	22	17	36	220	39
	Percent.	2.5%	1.9%	3.6%		6.1%	4.7%	10.0%	10.7%
Contact person	Freq.	24	18	20	47	49	39	139	29
	Percent.	6.6%	4.9%	5.5%	12.9%	13.4%	10.7%	38.1%	8.0%
Purpose of data collection	Freq.	7	4	8	12	33	68	213	19
	Percent.	1.9%	1.1%	2.2%	3.3%	9.1%	18.7%	58.5%	5.2%
Collection method	Freq.	10	12	8	23	41	70	192	8
	Percent.	2.75%	3.3%	2.2%	6.3%	11.3%	19.2%	52.8%	2.2%
Sample	Freq.	10	13	13	22	43	63	164	37
	Percent.	2.7%	3.6%	3.6%	6.0%	11.8%	17.3%	45.0%	10.1%
Weighting	Freq.	15	15	17	33	43	53	114	74
	Percent.	4.1%	4.1%	4.7%	9.1%	11.8%	14.6%	31.3%	20.3%
Response rates	Freq.	14	18	20	28	46	45	132	60
	Percent.	3.9%	5.0%	5.5%	7.7%	12.7%	12.4%	36.4%	16.5%
Bibliography	Freq.	21	21	21	34	45	43	109	65
	Percent.	5.9%	5.9%	5.9%	9.5%	12.5%	12.0%	30.4%	18.1%
Variables	Freq.	7	11	11	16	47	70	189	12
	Percent.	1.9%	3.0%	3.0%	4.4%	13.0%	19.3%	52.1%	3.3%
Question text	Freq.	8	9	12	15	31	54	177	55
	Percent.	2.2%	2.5%	3.3%	4.2%	8.6%	15.0%	49.0%	15.2%
Recoded and derived variables	Freq.	12	16	30	24	49	68	113	51
	Percent.	3.3%	4.4%	8.3%	6.6%	13.5%	18.7%	31.1%	14.1%

Frequencies of variables	Freq.	17	15	22	20	27	54	147	57
	Percent.	4.7%	4.2%	6.1%	5.6%	7.5%	15.0%	41.0%	15.9%
Data file formats	Freq.	13	13	9	31	19	57	198	20
	Percent.	3.6%	3.6%	2.5%	8.6%	5.3%	15.8%	55.0%	5.6%
Missing data	Freq.	21	21	37	43	42	50	121	24
	Percent.	5.9%	5.9%	10.3%	12.0%	11.7%	13.9%	33.7%	6.7%
Data collection instruments	Freq.	8	15	12	26	37	55	166	40
	Percent.	2.2%	4.2%	3.3%	7.2%	10.3%	15.3%	46.2%	11.1%

Table 28. Means and standard deviations of the 19 items for completeness

	Mean	Standard Deviation		Mean	Standard Deviation
Title	6.3	1.8	Bibliography	4.2	2.7
Principal investigator	4.9	2.8	Variables	5.8	1.8
Time period	6.3	1.5	Question text	5.1	2.6
Geographic location	5.9	2.0	Recoded and derived variables	4.6	2.5
Funding agency	5.5	2.4	Frequencies of variables	4.7	2.7
Contact person	4.8	2.3	Data file formats	5.6	2.1
Purpose of data collection	5.9	1.9	Missing data	4.7	2.3
Collection method	5.8	1.8	Data collection instruments	5.2	2.4
Sample	5.2	2.3	Response rates	4.5	2.6
Weighting	4.3	2.7			

Table 29. Correlation matrix for the items of absorptive capacity

	A. Professional Status	B	C.	D.	E.
B. Familiarity with topics of data	0.12**				
C. Experience in using the same data	0.23*	0.20*			
D. Experience using secondary data	0.31*	0.29*	0.26*		
E. Experience analyzing self-collected data	0.37*	0.14**	0.15**	0.31*	
F. Experience in collecting data	0.28*	0.14**	0.09***	0.25*	0.83*

*: p=0.00; **: p<0.05; ***: p<0.10.

Table 30. Value distributions of four items for absorptive capacity

		1	2	3	4	5	6	7
Familiarity with the topics of the data	Freq.	2	7	5	17	54	56	214
	Percent.	0.6%	2.0%	1.4%	4.8%	15.2%	15.8%	60.3%
Experience using the same data	Freq.	101	34	12	12	20	29	133
	Percent.	29.6%	10.0%	3.5%	3.5%	5.9%	8.5%	39.0%
Experience in secondary data analysis	Freq.	15	8	5	17	16	38	240
	Percent.	4.4%	2.4%	1.5%	5.0%	4.7%	11.2%	70.8%
Experience analyzing self-collected data	Freq.	48	13	14	15	24	34	186
	Percent.	14.4%	3.9%	4.2%	4.5%	7.2%	10.2%	55.7%
Experience collecting research data	Freq.	38	11	16	13	24	30	200
	Percent.	11.5%	3.3%	4.8%	3.9%	7.2%	9.0%	60.2%

Table 31. Means and standard deviations of four items for absorptive capacity

	Mean	Standard Deviation
Familiarity with the topics of the data	6.21	1.22
Experience using the same data	4.28	2.64
Experience in secondary data analysis	6.20	1.59
Experience analyzing self-collected data	5.40	2.25
Experience collecting research data	5.60	2.12

Table 32: Difference in overall quality between students and professors

	Students (doctoral, masters and undergraduates)		Professors (full, associate and assistant)	
	Frequency	Percentage	Frequency	Percentage
1	-	-	6	2.8%
2	2	2.3%	6	2.8%
3	6	6.8%	9	4.1%
4	9	10.1%	15	6.9%
5	21	23.6%	29	13.3%
6	23	25.8%	50	23.0%
7	28	31.5%	103	47.3%
Total	89		218	
Average ratings	5.6		5.8	

Table 33. Difference in sufficiency between students and professors

	Students (doctoral, masters and undergraduates)		Professors (full, associate and assistant)	
	Frequency	Percentage	Frequency	Percentage
1	-	-	1	0.5%
2	2	2.2%	9	4.0%
3	11	12.1%	10	4.5%
4	17	18.7%	25	11.2%
5	22	24.2%	48	21.5%
6	24	26.4%	66	29.6%
7	15	16.5%	64	28.7%
Total	91			
Average ratings	5.1		5.5	

Table 34. Difference in ease-of-use between students and professors

	Students (doctoral, masters and undergraduates)		Professors (full, associate and assistant)	
	Frequency	Percentage	Frequency	Percentage
1	1	1.1%	5	2.3%
2	4	4.5%	7	3.2%
3	4	4.5%	19	8.7%
4	11	12.4%	11	5.1%
5	21	23.6%	41	18.8%
6	29	32.6%	60	27.5%
7	19	21.4%	75	34.4%
Total	89		218	
Average ratings	5.4		5.6	

Table 35. Difference in ease-of-use between users familiar with the topics of data and users not familiar with the topics of data

	Familiar with the topics of data (rating 1, 2 and 3)		Familiar with the topics of data (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	2	14.3%	3	0.9%
2	3	21.4%	13	4.1%
3	2	14.3%	22	6.9%
4	-		25	7.9%
5	1	7.1%	66	20.8%
6	5	35.7%	91	28.6%
7	1	7.1%	98	30.8%
Total	14		318	
Average ratings	4.0		5.5	

Table 36. Difference in sufficiency between users familiar with the topics of data and users not familiar with the topics of data

	Familiar with the topics of data (rating 1, 2 and 3)		Familiar with the topics of data (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	2	14.3%	1	0.3%
2	1	7.1%	11	3.4%
3	2	14.3%	21	6.5%
4	2	14.3%	42	13.0%
5	5	35.7%	69	21.4%
6	2	14.3%	95	29.5%
7	-	-	83	25.8%
Total	14		322	
Average ratings	3.9		5.4	

Table 37. Difference in overall quality between users familiar with the topics of data and users not familiar with the topics of data

	Familiar with the topics of data (rating 1, 2 and 3)		Familiar with the topics of data (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	3	21.4%	5	1.6%
2	1	7.1%	8	2.5%
3	3	21.4%	14	4.4%
4	1	7.1%	25	7.8%
5	1	7.1%	50	15.7%
6	4	28.6%	77	24.1%
7	1	7.1%	140	43.9%
Total	14		319	
Average ratings	3.9		5.8	

Table 38. Difference in ease-of-use between users experienced and users not experienced in using the same data

	Experience with the same data (rating 1, 2 and 3)		Experience with the same data (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	6	4.1%	-	-
2	10	6.9%	4	2.2%
3	11	7.6%	13	7.2%
4	15	10.3%	11	6.1%
5	34	23.5%	34	18.9%
6	38	26.2%	51	28.3%
7	31	21.4%	67	37.2%
Total	145		180	
Average ratings	5.1		5.8	

Table 39. Difference in sufficiency between users experienced and users non-experienced in using the same data

	Experience with the same data (rating 1, 2 and 3)		Experience with the same data (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	2	1.4%	-	-
2	8	5.4%	4	2.2%
3	17	11.6%	6	3.3%
4	27	18.4%	18	10.0%
5	36	24.5%	40	22.1%
6	35	23.8%	52	28.7%
7	22	15.0%	61	33.7%
Total	147		181	
Average ratings	4.9		5.7	

Table 40. Difference in overall quality between users experienced and users non-experienced in using the same data

	Experience with the same data (rating 1, 2 and 3)		Experience with the same data (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	5	3.4%	5	1.1%
2	6	4.1%	6	1.7%
3	12	8.2%	12	1.7%
4	11	7.5%	11	7.8%
5	24	16.4%	24	15.6%
6	38	26.0%	38	24.6%
7	50	34.3%	50	47.5%
Total	146		179	
Average ratings	5.4		6.0	

Table 41. Difference in ease-of-use between users who are experienced in secondary data analysis and users who are not

	Experience in secondary data analysis (rating 1, 2 and 3)		Experience in secondary data analysis (rating 1, 2 and 3)	
	Frequency	Percentage	Frequency	Percentage
1	1	3.6%	5	1.7%
2	5	17.9%	9	3.1%
3	3	10.7%	22	7.6%
4	1	3.6%	25	8.7%
5	6	21.4%	58	20.1%
6	9	32.1%	78	27.0%
7	3	10.7%	92	31.8%
Total	28		289	
Average ratings	4.6		5.5	

Table 42. Difference in sufficiency between users who are experienced in secondary data analysis and users who are not

	Experience in secondary data analysis (rating 1, 2 and 3)		Experience in secondary data analysis (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	2	7.1%	-	-
2	2	7.1%	10	3.4%
3	4	14.3%	18	6.2%
4	4	14.3%	41	14.0%
5	6	21.4%	63	21.6%
6	5	17.9%	84	28.8%
7	5	17.9%	76	26.0%
Total	28		292	
Average ratings	4.6		5.4	

Table 43. Difference in overall quality between users who are experienced in secondary data analysis and users who are not

	Experience in secondary data analysis (rating 1, 2 and 3)		Experience in secondary data analysis (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	2	7.1%	5	1.7%
2	3	10.7%	7	2.4%
3	3	10.7%	13	4.5%
4	1	3.6%	25	8.7%
5	5	17.9%	42	14.5%
6	4	14.3%	76	26.3%
7	10	35.7%	121	41.9%
Total	28		289	
Average ratings	5.0		5.8	

Table 44. Difference in ease-of-use between users experienced in collecting and analyzing primary data and users not experienced

	Experience with primary data (rating 1, 2 and 3)		Experience with primary data (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	1	1.5%	5	2.0%
2	3	4.4%	11	4.5%
3	7	10.1%	18	7.4%
4	7	10.1%	20	8.2%
5	14	20.3%	49	20.0%
6	23	33.3%	64	26.1%
7	14	20.3%	78	31.8%
Total	69		245	
Average ratings	5.2		5.5	

Table 45. Difference in sufficiency between users who are experienced in collecting and analyzing primary data and users who are not

	Experience with primary data (rating 1, 2 and 3)		Experience with primary data (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	2	2.9%	-	-
2	1	1.5%	11	4.4%
3	11	15.9%	11	4.4%
4	8	11.6%	37	14.9%
5	19	27.5%	55	22.2%
6	18	26.1%	67	27.0%
7	10	14.5%	67	27.0%
Total	69		248	
Average ratings	5.0		5.4	

Table 46. Difference in overall quality between users who are experienced in collecting and analyzing primary data and users who are not

	Experience with primary data (rating 1, 2 and 3)		Experience with primary data (rating 5, 6 and 7)	
	Frequency	Percentage	Frequency	Percentage
1	2	2.9%	5	2.0%
2	2	2.9%	8	3.3%
3	4	5.8%	10	4.1%
4	3	4.4%	24	9.8%
5	14	20.3%	36	14.7%
6	17	24.6%	61	24.9%
7	27	39.1%	101	41.2%
Total	69		245	
Average ratings	5.67		5.71	

APPENDIX B

Survey Instrument

This survey is about a dataset that you analyzed for research , where you are not involved in collecting the data.

1. What is the title of the dataset (collected by other people) you analyzed most recently? (If the dataset doesn't have a title, please make one that best describe the topics of the data.)

2. When did you use that dataset?

3. What do you use that data for?

- Research
- Teaching
- Other, please specify________

4. Where did you obtain that dataset?

5. In analyzing that dataset, did you collaborate with the people who collected the data?

- Yes
- No.

6. Who collected that dataset?

7. What kind of entity collected that dataset?

- An individual researcher
- A small research group (2-5 people)
- A large research group (6 and more people)
- A private research organization
- A government organization
- Don't know
- Other, please specify________

8. What was the type of that dataset? please choose all that apply.

- Survey/interview data
- Census data
- Observation data
- Experimental data
- Administrative records
- Qualitative data
- Quantitative data
- Longitudinal data
- Cross-sectional data
- Other (please specify)__________

9. Roughly when that dataset was collected? E. g. in the 1990s

10. Roughly how big was the sample size (n) of that dataset?

11. Roughly how many variables were in the dataset?

12. Roughly how many questions were asked in the original data collection instrument?

13. What is your research field?

14. At the time that you used the data, which of the followings best describes your professional status?

- Master student
- Doctoral student: pre-candidate
- Doctoral candidate
- Post-doc researcher
- Assistant professor
- Associate professor
- Full professor
- Other (please specify)_____________

15. What kinds of documentation did you get for the data? Please choose all that apply. Documentation refers to the materials accompanying the data you use. It provides information about the data and helps you understand the data. It usually includes codebooks, project reports, etc.

- Code book(s)
- Reports about the data collection project
- Data collection instruments, such as survey or Interview questions
- Related bibliography
- Other (please specify)______________

16. How well each of the items below was described in the documentation? Please rate on scales from 1-7. 7 means most sufficiently documented. Please write down N/A if it is not applicable.

- Title of the dataset
- Principal investigator(s) of the study
- Time period covered by the data
- Geographic location where the data were collected
- Funding agency/sponsor
- Contact person(s) who are responsible for answering questions about the data
- Purpose and goals of the data collection
- Data collection method
- Sample and sampling procedures
- Weighting information
- Response rates
-
- Bibliography of publications related to the data
- Variables
- Question text
- Recoded and derived variables
- Frequencies of variables
- Data file formats
- Missing data
- Data collection instrument(s)

17. How much do you agree with the following statements? 1 means strongly disagree, 7 means strongly agree. Please write down N/A if it is not applicable.

- The topics of the data were within my specialty.
- The data was well documented.
- The documentation provided enough information for me to judge the reliability of the data.
- Learning to use the documentation was easy for me.
- It was easy to find the information I needed from the documentation.
- The documentation provided enough information for my purpose of use.

- With the documentation, I did not need additional information about the data for my purpose of use.
- The content of the documentation was clear and understandable.
- Overall, the documentation of the data was easy to use.
- With the documentation, I did not need additional information about the data for my use.
- I had prior experiences analyzing exactly the same dataset.
- I had prior experiences analyzing data collected by other people.
- I had prior experiences analyzing data collected by myself.
- I had prior experiences collecting research data.
- The documentation provided enough information for my purpose of use.

18. Besides using the documentation of the data, where else did you obtain information to help you understand the data? Please choose all that apply.

- Not applicable, I only used the documentation of the data.
- Data producers (people who collected the data).
- Other secondary users of the same data.
- Data archivists
- Publications based on the data.
- Websites of data producers
- Websites of data archives.
- Workshop
- Other (please specify)

19. How did you get information from other people (data producers, archivists, or other secondary data users, etc)? Please choose all that apply.

- Not applicable. I did not obtained any information about the data from other people.
- Email or telephone.
- Face-to-face conversations.
- I worked together with them for a period of time

- Other (please specify)___________

20. Why did you choose other channels to obtain information about the data, rather than obtain that information from the documentation of the data? choose all that apply.

- The documentation does not contain that information.
- The documentation is hard to use.
- Other information sources or channels are immediately accessible.
- I happen to encounter that information about the data.
- That information is tacit knowledge that is hard to be documented.
- Other (please specify)________________

21. If you think the documentation was hard to use, please explain why? otherwise, skip this question.

22. Can you give some examples of the information you obtained from outside of the documentation?

23. How much do you agree with the following statement? 1 means strongly disagree, 7 means strongly agree.

- I obtained good results from using the data.

24. Any additional comments about the documentation and data you used?

25. At the time that you used that dataset, what was your highest degree?

26. In what year did you obtain your highest degree?

27. What is your age?

28. What is your gender?

- Male

- Female

29. At the time that you used that dataset, which organization were you affiliated with?

30. At the end of my research project, how do you want the survey data collected today to be disposed?

- I want it to be destroyed.
- After all my identification information is removed, it can be deposited into a data archive and shared with the public.
- Other (please specify)____________

APPENDIX C

Interview Invitation Letter

Dear XXX,

I am a doctoral candidate at the School of Information, University of Michigan. My dissertation is about the secondary use of social science data. In my study, secondary use is defined this way: a user used a data set, and the user is not involved in the collection process of the data. I am trying to identify factors affecting the use of social science data, especially the difficulties that users encounter, the causes of the difficulties, and how users overcome the difficulties. The results of this study are expected to inform data sharing policies, improve services of data archive and help users using secondary data.

I am writing to you because I am informed that you have the experiences of using social science data. If you agree to participate in my study, I will schedule an interview with you, which I anticipate will last about 1 hour. The interview will be held at a time that is convenient for you; and with your permission, it will be audiotaped. All information you provide will remain confidential. I will ask you questions about your experiences in using secondary data.

Thank you for considering my request. I hope gain your participation in my study. In the meantime, please feel free to contact me if you have any questions.

Sincerely
Jinfang Niu
niujf@umich.edu
Phone: 734-330-7264

APPENDIX D

Survey Invitation Email (to secondary data users)

Dear [CustomValue] [LastName],
I am a doctoral candidate at the School of Information, University of Michigan. I am writing to you because I am informed that you have experience analyzing data collected by other people, and I am doing my dissertation on that topic. More specifically, I am doing an online survey about a dataset that you analyzed for research, where you are not involved in collecting the data. The dataset can be qualitative or quantitative data, census data, survey/interview data, experiment data, or administrative records, etc. I am trying to find out to what extent the documentation of the data was sufficient, and what you did when the documentation was inadequate. Documentation refers to the materials accompanying the data you use. It provides information about the data and helps you understand the data. It usually includes codebooks, project reports, etc. Findings from this survey will be helpful in deciding how to improve documentation quality, help secondary data analysis and eventually facilitate the scientific progress of the society.

The survey takes about only 15 minutes. Here is a link to the survey:
http://www.surveymonkey.com/s.aspx

Thank you for considering my request. Please feel free to contact me if you have any questions.

Sincerely,
Jinfang Niu
School Of Information, University of Michigan
niujf@umich.edu

APPENDIX E

Survey Invitation Email (to data librarians)

Dear [CustomValue] [LastName],
I am a doctoral candidate at the School of Information, University of Michigan. I am writing to you because I am informed that you have connections with secondary data users. Secondary data users are people who use data for research, where they are not involved in collecting the data. The data can be qualitative or quantitative data, census data, survey/interview data, experiment data, or administrative records, etc.

I am doing an online survey about secondary data analysis, which is part of my dissertation project. I am trying to find out to what extent the documentation of data is sufficient, and what users do when documentation is inadequate. Documentation refers to the materials accompanying data. It provides information about the data and helps users understand data. It usually includes codebooks, project reports, etc. Findings from this survey will be helpful in deciding how to improve documentation quality, help secondary data analysis and eventually facilitate the scientific progress of the society.

I would appreciate greatly if you forward my email to secondary data users. The survey takes about only 15 minutes. Here is a link to the survey:
http://www.surveymonkey.com/s.aspx?sm=7ob70FCiu7rRfPqPX1LX9w_3d_3d

Please feel free to contact me if you have any questions.

Sincerely
Jinfang Niu
School Of Information, University of Michigan

APPENDIX F

Consent Form (survey)

Purpose of the study

This survey is about a dataset that you analyzed for research , where you are not involved in collecting the data. The dataset can be qualitative or quantitative data, survey/interview data, experiment data, or administrative records, etc. I am trying to find out to what extent the documentation of the data was sufficient for your purpose of use, and what you did when the documentation was inadequate. Documentation refers to the materials accompanying the data you use. It provides information about the data and helps you understand the data. It usually includes codebooks, project reports, etc. Findings of this study will be helpful in deciding how to improve documentation quality and help secondary data analysis.

What risks and benefits are associated with my participation?

I do not foresee any risks to you other than a possible breach of confidentiality. To protect against that risk, various measures would be taken. Access to data will be limited to myself only. All information collected will remain confidential except as may be required by federal, state, or local law. Your name will not appear in any publication or public statement based on the study, all names or other potential identifying information will be omitted or changed. How the data will be disposed upon the completion of this project totally depends on your choice at the end of the survey. The goal of the study is to help users use secondary data, improve the efficiency of data sharing, and eventually facilitate the scientific progress of the society. As a data user, you will benefit both directly and indirectly.

What are my rights as a respondent?

Your participation is voluntary. You may ask questions, both before agreeing to be involved and during the course of the study, and they will be answered fully. You may refuse to participate before the study begins, discontinue at any time, or skip any questions.

Should you have questions regarding this study or your participation, please contact Jinfang Niu, doctoral candidate at the School of Information, University of Michigan, email: niujf@umich.edu, phone: 734-330-7264.

If agree to participate, please click the “next ” button to take the survey. Thanks very much for your participation.

APPENDIX G

Consent Form (interviews)

The purpose of this research project is looking into the experiences of using social science data, trying to find the difficulties of using, the causes of the difficulties, and how users overcome the difficulties. Findings from this study would provide background information for policy makers in making data sharing policies, for the data archives to improve their services and documentation guidelines, and therefore help users use data.

Your participation is voluntary. You may ask questions, both before agreeing to be involved and during the course of the study, and they will be answered fully. You may refuse to participate before the study begins, discontinue at any time, or skip any questions. How the interview data from you will be used by me totally depends on your choice.

After the investigator finishes this research, how do you want the interview data collected today to be disposed?

- I want the interview with me be destroyed.
- After all my identification information is removed, the interview with me can be deposited into a data archive and shared with the public.
- The interview with me can be deposited into a data repository and made accessible in a controlled environment. Users need to officially sign for not disclosing my identification information to others.

Can I record this interview?

BIBLIOGRAPHY

Arthur, D. J. and Stevens, K. T. (1989). Assessing the Adequacy of Documentation Through Document Quality Indicators. In Proceedings of the Conference on Software maintenance, Miami, 16-19. Oct. Washington D.C.: IEEE Computer Society Press (pp. 40-49)

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., & Wouters, P. (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. Data science Journal, 3(29), (pp.135-152).

Birnholtz, J. and Bietz, M. (2000). Data at Work: Supporting Sharing in Science and Engineering. ACM conference.

Blommaert, J. (1997). Workshopping: Notes on Professional Vision in Discourse Analysis. Wilrijk: Antwerp Papers in Linguistics 91.

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

Borgman, C. L. (2007). Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, MA: MIT Press.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2006). Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. In: Proceedings of the Tenth European Conference on Digital libraries (Alicante, Spain), 170-183. Berlin, Springer.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries. Vol. 7, numbers 1-2. October.

Boruch, R. F. (1978, Ed.). Secondary analysis. New Directions for program Evaluation, no. 4. San Francisco: Jossey-Bass.

Boruch, R. F. (1985). Definitions, products, distinctions in data sharing. In S. E. Fienberg, M. E. Martin, & M. L. Straf (Eds.), Sharing research data (pp. 89-122). Washington, DC: National Academy Press.

Boruch, R. F., et al. (1991). Sharing Confidential and sensitive data. In Sieber, J. E. (Ed.), Sharing social science data: advantages and challenges. SAGE publications.

Bowering, D. J. (1984, Ed.). Secondary analysis of available data bases. San Francisco: Jossey-Bass.

Breusch, T. and Holloway, S. (2004). Australian social science data archive. The Australian economic review, vol. 37, no. 2, pp. 222-9.

Buckland, M. (1991). Information as thing. Journal of the American Society of Information Science 42:5 (351-360)

Campbell, E.G., Clarridge, B. R., Gokhale, M. L., Birenbaum, S., Hilgartner, Holtzman, N. A. & Blumenthal, D. (2002). Data withholding in academic genetics: Evidence from a national survey. *JAMA* 287(4): 473-480.

Carlson, S., and 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), article 15. http://jcmc.indiana.edu/vol12/issue2/carlson.html

Carmines, E. G., & Zeller, R. A. (1979). Reliability and Validity Assessment. Thousand Oaks, CA: Sage Publications.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. Cognition and Instruction 19(3), 323-393.

Clubb, M. J., Erik, W. A., Geda L. C., and Traugott, W. M. Sharing research data in the social sciences. In Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds.). (1985). Sharing research data. Washington, DC: National Academy Press.

Collins, H. M. (2001). Tacit Knowledge, Trust, and the Q of Sapphire' Social Studies of Science, 31, 1, 71-85

Corti, L. (2000, December). Progress and Problems of Preserving and Providing Access to Qualitative Data for Social Research - The International Picture of an Emerging Culture. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research [Online Journal], 1(3).

Corti, L. (2002). Qualitative data processing guidelines. Qualidata, UK Data Archive, University of Essex, Colchester.

Corti, L. (2005) Qualitative Archiving and Data Sharing: Extending the reach and impact of qualitative data, IASSIST Quarterly, 29(3).

Corti, L. & Bishop, L. (2005, February). Strategies in Teaching Secondary Analysis of Qualitative Data [67 paragraphs]. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [On-line Journal], 6(1), Art. 47. Available at: http://www.qualitative-research.net/fqs-texte/1-05/05-1-47-e.htm [Date of Access: Jan. 14, 2008].

Council on Governmental Relations. (2006). Access to and retention of research data: rights and responsibilities. http://206.151.87.67/docs/CompleteDRBooklet.htm

[Data of Access: Jan. 14, 2008].

Dale, A., Arber, S. and Procter, M. (1988). Doing Secondary analysis. London, Unwin Hyman.

David, M. (1991). The science of data sharing: Documentation. In Sieber, J. E. (Ed.), Sharing social science data: advantages and challenges. SAGE publications.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-340.

Dierickx, I. and Cool, K. (1989). Asset stock accumulation and the sustainability of competitive advantage: reply. Management Science. Vol. 35, No. 12.

ESRC (Economic and Social Research Council), (2000). Economic and Social Research Council data policy. http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/DataPolicy2000_tcm6-12051.pdf

Estrin, D., Michener, W. K. & Bonito, G. (2003). Environmental cyberinfrastructure needs for distributed sensor networks: A report from a National Science Foundation sponsored workshop. Scripps Institute of Oceanography. Visited http://www.lternet.edu/sensor_report/ on 12 May 2006.

Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds.). (1985). Sharing research data. Washington, DC: National Academy Press.

Frank, R. H. (2007) Externalities, Property Rights and the Coase Theorem. Chapter 16 of Microeconomics and Behavior, 6th ed. McGraw-Hill.

Grudin, J. (1994). Eight Challenges For Developers. Communications of the ACM. Vol. 37, No.1. January.

Gutmann, M. , K Schürer, D Donakowski and Hilary Beedham. (2004). The selection, appraisal, and retention of digital social science data. Data Science Journal, Volume 3, 30.

Halstuk, M. E. and Chamberlin, B. F. (2006). The Freedom of Information Act 1966-2006: A retrospective on the rise of privacy protection over the public interest in knowing what the government's up to. Communication law and policy [1081-1680] vol:11 issue: 4 pg: 511 -564

Hyman, H. H. (1987). Secondary analysis of sample surveys, with a new introduction. Wesleyan University Press. Middlettown, Connecticut.

ICPSR (Inter-university Consortium for political and social research). (2005). Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle. http://www.icpsr.umich.edu/access/dataprep.pdf

Ivari, J. and Linger, H. (1999). Knowledge work as collaborative work: a situated activity theory view. In: Proceedings of the Thirty-Second Annual Hawaii International Conference on Systems Sciences, IEEE Computer Society Press, Los Alamitos, CA.

Kiecolt, k. J., and Nathan, L. E., (1985). Secondary analysis of survey data. Sage Publications (Beverly Hills).

Jacobs, J. A. and Humphrey, C. (2004) Preserving research data. Communications of the ACM. Vol. 47, 9: 27–29.

Lehner, F. (1993) Quality control in software documentation based on measurement of text comprehension and text comprehensibility. Information Processing & Management, Volume 29, Issue 5, September-October, pp 551-568.

Lesk, M. (2004). Online Data and Scientific Progress: Content in Cyberinfrastructure. Presentation given as part of the UK Digital Curation Centre's Visitor Programme. Edinburgh: 24 September, 2004. [Available: http://www.dcc.ac.uk/docs/bl-sep04a.ppt.]

McCall, R. B. and Applebaum, M. I. (1991). Some issues of conducting secondary analysis. Developmental psychology. Vol. 27, No. 6, 911-917.

MacKenzie, D. and Spinardi, G. (1995). Tacit Knowledge, Weapons Design and the Uninvention of Nuclear Weapons, American Journal of Sociology, Vol. 101, No. 1, 44-99.

Markus, M. L. (2001) Toward a theory of knowledge reuse: type of knowledge reuse situations and factors in reuse success. Journal of Management Information Systems. 18(1), 57-93.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). Plans and the structure of behavior. New York: Holt, Rinehart & Wilson.

Moran, T. P., Chiu, P., Harrison, S., Kurtenbach, G., Minneman, S.; and Melle, W.V. (1996). Evolutionary engagement in an ongoing collaborative work process: A case study. In Proceedings of the ACM 1996 Conference on Computer-Supported Cooperative Work , Cambridge, MA, 150-159.

National Institutes of Health (2003). Data Sharing Policy and Implementation Guidance. Available: http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm.

National Institutes of Health. (2006). Data Sharing Policy and Implementation Guidance. Available: http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

National Research Council. (1997). Bits of Power: Issues, in Global Access to Scientific Data. Washington, D.C.: National Academies Press. http://www.nap.edu/catalog.php?record_id=5504

National Research Council. (2003). Sharing publication-related data and materials: responsibilities of producership in the life sciences, Washington, D.C.: National Academy Press.

Nelson, R. R. and Winter, S. (1982). An Evolutionary Theory of Economic Change. Harvard University Press, Cambridge, Massachusetts, and London, England.

Niu, J. and Hedstrom, M. (2007a). Incentives and barriers in data sharing ---- a survey report. Working paper. Not published.

Niu, J. and Hedstrom, M. (2007b). Streamlining the "Producer/Archive" Interface: Mechanisms to Reduce Delays In Ingest And Release Of Social Science Data. *DigCCurr2007 Conference*, Chapel Hill, NC (April 18-20).

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. Organizational Science. Vol. 5, No.1.

Nunally, J. C. (1978). Psychometric Theory (2 ed.). New York: McGraw-Hill.

Orlikowski, W.J. (1995). Evolving with Notes: Organizational change around groupware technology, CISR WP No. 279, Sloan WP No. 3823, CCS WP No. 186, Center for Information System Research, Sloan School of Management, MIT.

Polanyi, M. (1962). Personal knowledge: Toward a post-critical philosophy, Harper Torchbooks, New York.

Pritchard, S; Carver, L; Anand, S. (July 2004). Collaboration for Knowledge Management and Campus Informatics. Report to the Andrew W. Mellon Foundation" under award number 40200638.

Reidpath, D. D. & Allotey, P. A. (2001). Data Sharing in Medical Research: An Empirical Investigation, *Bioethics* Vol.15 (2): 125-134

Rogers, T. W., Anderson, J. O., Klinger, D. A. & Dawber, T. (2006). Pitfalls and potential of secondary data analysis of the Council of Ministers of Education, Canada, National assessment, Canadian Journal of Education 29, 3: 757-770

Sieber, J. E. (1991). Sharing social science data: advantages and challenges. Newbury Park, Calif: Sage Publications.

Strathern, M. (2005, March). Useful knowledge. Lecture presented at The Isaiah Berlin Lecture, Manchester, UK. Not published.

Szulanski, G. (1996). Exploring internal stickiness: Impediments to the transfer of best practice within the firm. Strategic Management Journal, 17(Winter Special Issue), 27-43.

Tausworthe, R. C. (1977). Standardized development of computer software, Prenice-Hall.

Tuomi, I. (1999). Corporate knowledge: Theory and Practice of Intelligent Organizations. University of Helsinki, Metaxis. Helsinki. 453 p. ISBN 951-98280-0-1.

Van den Berg, H. (2005, January). Reanalyzing Qualitative Interviews From Different Angles: The Risk of Decontextualization and Other Problems of Sharing Qualitative Data [48 paragraphs]. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [On-line Journal], 6(1), Art. 30. Available at: http://www.qualitative-research.net/fqs-texte/1-05/05-1-30-e.htm [Date of Access: Jan. 14, 2008].

Van House, N. A., Butler, M. H. and Schiff, L. R. (1998). Cooperative knowledge work and practices of trust: Sharing environmental planning data sets. in (eds.) Proceedings of CSCW 1998. ACM Press, New York.

Wicherts, J. M., Borsboom, D., Kats, J. and Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. American psychologist. Vol. 61(7) 726-728.

Wilson, T. D. (2002). The nonsense of ‘knowledge management’. Informational Research, Vol. 8 No. 1, October.

Wilson, R. E. and Maxwell, C. D. (2006, Oct). NIJ's Data Resources Program and the NACJD Paper presented at the annual meeting of the American Society of Criminology (ASC), Los Angeles Convention Center, Los Angeles, CA 2006-10-05. http://www.allacademic.com/meta/p143510_index.html

Zimmerman, A. (2003). Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists. Unpublished Dissertation, Information and Library Studies, University of Michigan, Ann Arbor.