Documentation Evaluation Model for Social Science Data

Jinfang Niu niujf@umich.edu

Margaret Hedstrom hedstrom@umich.edu

School of Information, University of Michigan.

SI North, 1075 Beal Avenue, Ann Arbor, MI 48109-2112

Abstract

Information technology and data sharing policies have made more and more social science data available for secondary analysis. In secondary data analysis, documentation plays a critical role in transferring knowledge about data from data producers to secondary users. Despite its importance, documentation of social science data has rarely been the focus of existing studies. In this paper, based on an introduction of the concept of documentation and its role in secondary data analysis, the authors proposed the Documentation Evaluation Model(DEM) for social science data. In the model, two indicators are used to evaluate the documentation for social science data: sufficiency and ease-of-use. Then the authors review the sufficiency problems of documentation, identify three factors that affect the sufficiency of documentation: users, data, and the ease-of-use of documentation, and formulate hypotheses about how those factors affect the sufficiency of documentation. In future research, a survey instrument will be created based on the model and the factors affecting the sufficiency of documentation. The survey instrument will then be applied to the secondary users of social science data. Hypotheses will be tested based on the survey data.

Introduction

Information technology has made large-scale data collection and long distance data sharing easier. Data repositories have been proposed as an important component of the cyber-infrastructure for scientific research (Atkins, Droegemeier, Feldman, Garcia-Molina, Klein, Messina, Messerschmitt, Ostriker & Wright, 2003). In the meantime, policy support for data sharing has become stronger. Since the implementation of the Freedom Of Information Act (FOIA) in 1966, administrative records collected by government agencies become 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 have started to require data sharing. The National Science Foundation (NSF) expects funded researchers to share data, samples, physical collections and other supporting materials under NSF awards (http://www.nsf.gov/pubs/gpg/nsf04_23/6.jsp). The National Institute of Justice (NIJ) requires all of its grantees to deposit data to a data archive after their research projects are completed (Wilson & Maxwell, 2006). According to the data sharing policy of National Institute of Health (NIH), investigators of grants that are above $500,000 in direct costs in any single year are expected to include a data sharing plan or state why data sharing is not possible (NIH, 2003). Data sharing is causing profound changes in scientific research. As Lesk (2004) asserts, the scientific paradigm has now 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). However, sharing of social science data, on which many publications and scientific findings are based, has received inadequate attention from the information science community. Only very few information scientists have conducted theoretical or empirical explorations on the secondary analysis of social science data.

Secondary data analysis is the analysis of data produced by someone else. More specifically, the users are not involved in the production process of the data directly (produce the data themselves) or indirectly (being part of the team who design the study and produce the data). In order to enable secondary data analysis, knowledge about the data needs to be transferred from the data producers to secondary users. Documentation accompanying data is one channel to transfer knowledge about the data. When documentation does not provide enough information about the data, users need to seek more information about the data from other channels. Getting more information from data producers by email, telephone, personal visiting or even by working together with data producers were reported as helpful (Birnholtz and Bietz, 2000; McCall and Applebaum, 1991; Zimmerman, 2003). Assistance from data producers, however, is considered as hard to get. Many data producers collect data for their own research. There is evidence that they do not want to share their data, or provide documentation for secondary use (Campbell et al., 2002; Reidpath and Allotey, 2001; Blumenthal, et al., 2006). We can imagine that providing personal instruction to secondary users would be even more demanding for them, especially for publicly available datasets which are very likely to be used by strangers. When data producers' help is not available, reading previously published literature using the data, browsing the websites of data producers, or consulting other knowing experts, including other secondary data users or data archivists were reported as very useful. (McCall and Applebaum, 1991; Sieber, 1991). For very complex data sets that are produced to be shared, such as the National Comorbidity Survey, attending a workshop to learn about the data is recommended for secondary users (http://www.hcp.med.harvard.edu/ncs/ncs_data.php).

Documentation plays a critical role in secondary data analysis. When documentation provides sufficient information about data, users do not need to seek other channels for more information, and data producers would not be bothered to provide personal help. Mass sharing and mass secondary data use are only possible when documentation is good enough to rule out deep involvement of data producers in secondary data analysis. Thus, to facilitate data sharing, an important step is to improve the quality of documentation. However, even though it was pointed out that inadequate documentation causes difficulties in secondary data analysis, documentation has rarely been the focus of existing studies on data sharing and secondary data analysis. In this paper, based on an introduction of the concept of documentation and its role in secondary data analysis, we will propose the Documentation Evaluation Model(DEM) for social science data, identify factors that affect the sufficiency of documentation, and formulate hypotheses about how the sufficiency of documentation vary with those factors.

Documentation

Documentation refers to the materials accompanying data that provides information about the data. Documentation is helpful for users to understand and analyze data, but the content and quality of documentation vary from one dataset to another. Documentation usually includes codebooks, and sometimes also includes related bibliographies and data collection instruments. Documentation sometimes is 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. In the context of social science data sharing, a Data Documentation Initiative (DDI) record is an example of structured metadata. Data collection instruments and texts of interview or survey questions are examples of unstructured metadata. According to Gutmann, Schürer, Donakowski and Beedham (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 variables or groups of variables. (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 social science data community. "Metadata" is used in the secondary use of natural science data, such as ecology, physics and biology data etc.

The formats of documentation vary from penciled notes, digital Word or PDF files, or even web pages. 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 mails or emails. Documentation for some very large data sets is published online as hyperlinked web pages. For example, http://www.icpsr.umich.edu/CPES/index.html is the website for the National Institute Health Collaborative Psychiatric Epidemiology Surveys (CPES).

The role of documentation in secondary data analysis

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. Some of that metadata information is often extracted from documentation and put into the 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 researcher 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 numbers 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 and or incorrect conclusions might be drawn from the data.

Documentation Evaluation Model (DEM) for social science data

Here we propose the Documentation Evaluation Model(DEM) for social science data. The purpose of the model is to measure how sufficient and easy-to-use documentation of social science data is for users. Therefore the model is proposed in a way that evaluation can be conducted based on user reported information. This premise decides that only subjective measures can be selected for this model, and that different users might evaluate the same documentation differently.

DEM evaluates documentation of social science data on two dimensions: sufficiency and ease-of-use. Sufficiency means the extent to which the documentation provides enough information about the data so that users do not need to seek information about the data from other channels. One indicator for sufficiency is completeness, which is adapted from the Document Quality Indicators (DQI) for software documentation (Arthur and Stevens, 1989). Completeness in DQI means all the required components are present in the software documentation. This indicator is adapted in two ways. First, instead of using the required components for software documentation, the required components for DEM are decided based on DDI. Second, instead of asking users to check whether each required component "exists or not", we ask users to evaluate how well each required component is described in documentation. This will convey more information than the binary value of "exists or not". Completeness does capture much of the meaning of sufficiency, but we do not believe that users necessarily consider the documentation sufficient if only each required component is well described. Therefore, besides completeness, we added indicators for users' general perceptions of documentation sufficiency.

To sum up, sufficiency is measured by two groups of indicators:

- Completeness. The completeness of documentation is measured by how well each required component is described in the documentation. Below is a list of required components of good documentation. The list was created based on the international standard DDI.

  Title, principal investigator, time period covered by the data, geographic location where the data were collected, funding agency/sponsor, contact persons who are responsible for answering questions about the data, purposes and goals of the data collection, data collection method, sample and sampling procedures, weighting information, response rates, bibliography of publications related to the data, variable labels, value labels, question text, recoded and derived variables, frequencies of variables, data file type, missing data, data collection instruments.

- General perception of sufficiency: users' overall perception of the extent to which the documentation provides enough information about the data for their purpose of use.

Ease-of-use means the extent to which the documentation is perceived by the users as easy to use.
The measures for this construct are adapted from the Technology Acceptance Model (TAM) (Davis, 1989). TAM uses two subjective indicators to predict how likely technology products could be accepted by customers (Davis, 1989). The two indicators are: perceived ease-of-use and perceived usefulness. 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. Those measures are tailored for documentation of social science data. Since documentation is not operable, measure "b" and "e" is deleted. Measure "c" is 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. Measures for perceived usefulness are based on how much better off the user is using the product versus not using the product, such as time saving, increased productivity, enhanced effectiveness and effort reduction. Since using data with documentation is almost undoubtedly better off than using data without documentation. So I did not adopt "perceived usefulness" in DEM. To sum up, measures for ease-of-use are:

- How easy it is to learn to use the documentation.
- How easy it is to find information from the documentation.
- How clear and understandable the content of the documentation is.

Problems with documentation

Despite its important role, documentation is commonly reported as problematic (Fienberg, Martin and Straf, 1985; Club, Austin, Geda and Traugott, 1985; Van Den Berg, 2005; Corti, 2000; McCall and Applebaum, 1991; Zimmerman, 2003; Borgman, 2007). The problems can be grouped into two categories. The first problem is called inherent insufficiency. This problem is attributed to the nature of communication and tacit knowledge. According to the theory of communication reductionism, not everything can, or should be transferred. Some kind of reduction, and thus loss of complexity is inevitable (Strathern, 2005; Carlson and Anderson, 2007). Some tacit knowledge is difficult to articulate and therefore cannot be documented. When undocumentable tacit knowledge is important for secondary data analysis, documentation is inherently insufficient (Birnholtz and Bietz, 2000). In this case, improving the sufficiency of documentation would be very hard, the cost of which may overtake the cost for users seeking other channels for information. This problem reveals the inherent limitation of documentation as a channel to transfer knowledge from data producers to secondary data users. The solution of this problems is less about improving the quality of documentation, but to utilize other channels and sources to supplement documentation, such as reading previously published literature using the data, browsing the websites of data producers, consulting data producers or other knowing experts, including other secondary data users or data archivists.

The second type of problem is poor documentation, which means the documentation is insufficient and/or not easy to use because the data producers did not do an adequate job preparing documentation. Data producers do not document the data well for secondary users either because they are not motivated to do so, or because they lack the knowledge or skills to document for secondary users or both. Many data producers do not want to share their data for various reasons. When data is shared with the public, data producers lose the exclusive rights to publish follow-up papers based on the data. The data may no longer be used as barter in exchange for others' data or for funds, equipment, and other resources (Borgman, 2007, p174). Some data producers also worry about the misuse and misinterpretation of data by unqualified users, and the possibility of being accused of misconduct etc. Few institutions have formal policies and procedures for access to and retention of research data (Council on Governmental Relations, 2006). Even though some funding agencies do have data sharing policies, there is 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 data from prior grants were made available (Council on governmental relations, 2006.) Some data sharing policies forbid some of the direct benefits researchers might obtain from sharing their data. For example, NIH and the National Research Council (2003) claimed that it is not acceptable for the data provider to require collaboration or co-authorship of future publications as a condition of providing data. The cost of data sharing is also a disincentive. To make data usable by secondary users, data producers often need to prepare their data. Data preparation includes checking the integrity and consistency of data, careful naming of variables and choosing variable labels, organizing the variables, such as grouping them to enable secondary analysts to get an overview of the data quickly, etc. When human subjects are involved in a study, ensuring anonymity of the subjects before sharing is critical. All the direct identifiers, such as name, address, telephone numbers, and Social Security Numbers, indirect identifiers[i] and other information that could lead to "deductive disclosure" of participants' identities should be removed or processed in a way that doesn't hurt the value for secondary analysis. (ICPSR, 2005).

If data producers do not have incentives to share data in the first place, it seems safe to assume that they would be unwilling to spend effort to document data well for secondary data users. Even if they have to share data according to FOIA or data sharing policies of funding agencies, they may "comply with the letter of the law rather than its spirit, depositing poorly documented data that of little value. " (Borgman, 2007; Borgman, Wallis, Enyedy, 2007)

One might argue that data producers need to manage and document data for their own research anyway, and therefore documentation would not require extra effort from data producers. But documenting data sufficiently for others requires considerably more time and effort than documenting them only for the use of a small research team (Borgman, Wallis, Enyedy, 2007). Documenting data for later use also requires much more effort than what is required to publish data summaries in a journal article or conference papers (Borgman, Wallis, Enyedy, 2007). Documentation for self-use tends to be incomplete and informal. Investigators may be able to accomplish their own research goals without fully cleaned and well-documented data. Some researchers keep the details of data collection, variable construction, and particular quirks of the data in their own memory and do not put them in writing (Fienberg, Martin & Straf, 1985; Breusch and Holloway, 2004). 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 only for self-use do not have to be technically accurate and correct (Orlikowski, 1995), they often contains shorthand detail that makes sense only to the author (Markus, 2001). As Ackerman (1994) pointed out: "the shorter the distance the information might travel or the less likely it was that the information could be viewed by strangers, the more informal the information content was likely to be." Documentation for self-use tends to be useful to the author in the short term. Sometimes those documents are even hard to use by the producers themselves after a period of time, let alone by secondary users (Moran, et a. 1996).

When people have purposely created documentation for self-use, they may strenuously resist making the documentation public, or they may need to take extra effort to re-shape the documentation for secondary use. The effort required to explain one's research adequately increases as a function of the distance between data producers and users. Documenting data for use by team members is more difficult than documenting them for self-use. Documenting them for off-site collaborators is still more difficult. The most difficult of all is documenting for unknown future users (Borgman, 2007, p167; Markus, 2001), which is precisely the case in public data sharing.

The problem of poor documentation could be solved or alleviated by providing appropriate incentives and instruction to researchers who are required to share data. Some funding agencies provide financial support for the cost involved in data sharing. 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). Even today, a survey on NIJ grantees found that many grantees expected more financial support for documentation (Niu & Hedstrom, 2007). (Arzberger, et al. 2004) pointed out scant attention to data management happens in many areas of public research. NIH data sharers are allowed to charge data requestors for the costs associated with sharing data (NIH, 2003). There is no empirical study showing how effective that mechanism is for improving documentation quality.

Data archives share the cost of data preparation by processing and improving the quality of data and documentation. But their work has to build on the work done by data producers. Data producers should take the main responsibility for documenting data for sharing (Zimmerman, 2003).

Data archives have tried to instruct data producers to prepare documentation. The data archive community has created a documentation standard specially for social science data. It is the international standard DDI (ICPSR, 2005). DDI gives a comprehensive list of items and elements that should be provided with the data for secondary use. Documentation following DDI is supposed to help data producers prepare data for secondary use. Unfortunately, according to a 2006 survey on NIJ grantees who are required to deposit data to a data archive, 68% of the respondents don't know what DDI is. Of those who know DDI, 29% don't know how to use it. 12% think it make the documentation work even more complicated. Only 17% gave positive response and said it reminds them of the items they need to provide (Niu & Hedstrom, 2007). Some data archives like ICPSR provide detailed guidelines about how to prepare data for deposit. According to the same 2006 survey, the data documentation guidelines provided by ICPSR are considered useful, but many depositors are not aware of the data documentation guidelines (Niu & Hestrom, 2007).

Data archives also proposed the idea of starting to document data early in the research process. According to this idea, a data sharing and archiving plan should be fleshed out while the researcher is at the stage of outlining and writing the actual grant application (Jacobs and Humphrey, 2004; ICPSR, 2005). Important issues can be addressed by thinking ahead during the early phase of the project. For example, if the data is going to be shared with the public, it might be critical to include that information in the informed consent to human subjects, to determine the cost of preparing and archiving data early, so it is possible to apply funding for that cost, to identify potential users early, so it is possible to cater for users' needs when documenting data. By planning early and incorporating data preparation requirements into the life cycle of the research project, researchers could keep in mind the requirements, to consciously or unconsciously shape their documentation for secondary use. With the early documentation idea implemented, documentation for secondary use will no longer be an additional burden at the end of research projects. Rather it becomes a byproduct of research projects. It is supposed to alleviate the cost concern of data producers. However, how the early documentation concept has been implemented in practice is not very clear. The 2006 survey found that data producers still complained about the problems caused by documenting at the end of research projects. (Niu & Hestrom, 2007)

Factors affecting sufficiency

What we have learned from the previous sections are: documentation is important, but it is often insufficient for secondary use. Funding agencies and data archives have taken some measures to improve the quality of documentation, but it is not clear how effective those measures are. Therefore we do not know the general situation of the sufficiency of documentation. Nevertheless, existing literature does discuss the characteristics of various kinds of data, incentives for data producers and other related issues. Based on the literature, we identified three factors affecting the sufficiency of documentation: the user, the data and the ease-of-use of documentation. Data vary in their production methods, time of production, distributors and producers. Below are some hypotheses of how these factors affect the sufficiency of documentation.

Users: users who are more familiar with the data or have shorter knowledge distance from data producers can infer more tacit knowledge and have less need to rely on documentation. Therefore controlling for the other factors, documentation should be perceived to be more sufficient by those users.

Qualitative vs. 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 (Van Den Berg, 2005; Corti, 2002). Knowledge about qualitative data is highly contextual and experience-dependent. Data producers gain this knowledge through their direct experience with 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 very difficult to articulate. Some researchers are concerned that qualitative data cannot be used sensibly without the accumulated background knowledge which 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. Therefore, documentation for qualitative social science data may be perceived to be less sufficient than that of quantitative social science data, controlling for all other factors.

Survey data vs. other kinds of data: Zimmerman (2003) found that an adequate description of methods suffice to convince ecologists of data quality when the data are simple and easy to collect. Stream survey data was used as an example of simple data. Boruch (1991, p80) also mentioned that survey data are simpler than field experiment data. Therefore documentation of survey data might be perceived to be more sufficient than documentation of data collected using other methods, controlling for the other factors.

Interview vs. conversational data: Qualitative interviews are generally considered to be more suitable for secondary analysis than conversational data. The analysis of conversations requires much more in-depth information about the history of the social relations between the conversational partners than does the analysis of interview discourse (Van Den Berg, 2005). Therefore documentation of interview data may be perceived to be more sufficient than that of conversational data, controlling for the other factors.

New data vs. old data: 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). Therefore, documentation of older data may be perceived to be less sufficient than that of newer data, controlling for other factors.

Data producers: Data producers can be individual researchers, small groups of researchers, government agencies, professional data producers such as the Institute of Social Research (ISR), the Research Triangle Institute (RTI) or the Census Bureau. When the data producer is a large group or organization, including professional data producers and government agencies, it is often the case that different individuals desig the study, those collect the data, process the data, and analyze data. Thus, in order to enable other people in the producer organization to understand the data, each team member has to make explicit the ways he proceeded. The larger the data producer group, the more loosely coupled the research 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. Government agencies often contract data collection to professional data producers. Professional data producers are experts in documenting data for secondary use. They tend to collect data to be shared with many users, which also requires high quality documentation. In cases where a single researcher collected and also interpreted the 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. Therefore, documentation produced by a single researcher is likely to be perceived to be less sufficient than that produced by large research groups, controlling for the other factors.

Distributor: Data might be distributed to users by data producers themselves or by data archives. Data archives normally process and enhance the quality of data and documentation before distribution. For documentation of low quality, data archives can make great improvements. On the contrary, for documentation that already is of high quality, there may not be much room for improvement. A hypothesis mentioned above is that the sufficiency of documentation is likely to vary across different data producers. If this is true, the effect of the distributor should be moderated by data producers. In other words, for data produced by a single person or by a small research group, there should be a bigger difference between the documentation distributed by the data producers and that distributed by data archives. For data produced by other kinds of producers, there should be less or no difference between the documentation distributed by the data producers and that distributed by data archives.

Ease-of-use: The ease-of-use of documentation has been found to affect users' information seeking behavior. According to data archivists, typical questions that users raise sometimes are already answered within the documentation accompanying the data. In some cases this is the result of the poor writing quality of the documentation (Siber, 1991). The change of information seeking behavior might affect user's perception of the sufficiency of documentation.

FUTURE RESEARCH

DEM and the factors affecting sufficiency will be operationalized. A survey questionnaire will be created based on the operationalizable variables and measurements. Then the questionnaire will then be applied to secondary users of social science data. The survey data will be used to test the hypotheses proposed above.

ACKNOWLEDGMENTS

This research is funded by the National Science Foundation, Award # IIS 0456022 as part of the project entitled "Incentives for Data Producers to Create Archive-Ready Data Sets."

We greatly appreciate the comments from Ricardo L. Punzalan.

References

Ackerman, M.S. (1994). Definitional and contextual issues in organizational and group memories, available from the author's website, http://www.ics.uci.edu/~ackerman.

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, Vol. 3, 29 November 2004. 135-152.

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messina, P., Messerschmitt, D. G., Ostriker, J. P., and Wright, M. H. (2003). Revolutionalizing science and engineering through cyber-infrastructure: Report of the National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure. National Science Foundation. http://www.nsf.gov/cise/sci/reports/atkins.pdf (accessed September 18, 2006)

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

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

Borgman, C.L., Wallis, J.C., Enyedy, N. (forthcoming). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries.

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

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

Boruch, R. F., et al. (1991). Sharing Confidential and sensitive data. In Siber, 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.

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

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.

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). Qualilitative 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.

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

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

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

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 iss: 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

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

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.]

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.

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

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.

Morkes, J. and Nielsen, J. (1998). Applying Writing Guidelines to Web Pages. Sun Microsystems. Internet. <http://www.useit.com/papers/webwriting/rewriting.html>

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

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

Niu, J. and Hedstrom, M. (2007). Incentives and barriers in data sharing ---- a survey report. Working paper.

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.

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.

Tsakonas, G. and Papatheodorou, C. (2006). Analyzing and evaluating usefulness and usability in electronic information services. Journal of Information Science. 2006; 32, 400.

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].

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 from 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.

[i] Indirect identifiers are variables that will not reveal identify of human subjects when used alone, but may do so when used in combination with other variables. For example, ZIP code may not be troublesome in the univariate case, but when combined with other attributes like race and annual income, a ZIP code may allow unique individuals (extremely wealthy, poor) residents of that ZIP code to become visible. If protecting confidentiality of human subjects diminishes ithe value of data for secondary analysis, data depositors could apply for restricted use.